TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

Department of STATISTICS
Statistical Interdisciplinary
Research Laboratory
E4I5EP@TAMVM1.BITNET

Emanuel Parzen
Distinguished Professor

BOUNDARY KERNEL ESTIMATION OF THE TWO SAMPLE COMPARISON DENSITY FUNCTION

Technical Report #56

May 1989

William Pyle Alexander

Texas A&M Research Foundation

Project No. 5641

Sponsored by the U. S. Army Research Office

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO 23010.5-MA	2. GOVT ACCESSION NO. N/A	3. RECIPIENT'S CATALOG NUMBER N/A
4. TITLE *(and Subtitle)* Boundary Kernel Estimation of the Two Sample Comparison Density Function.		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) William Pyle Alexander		8. CONTRACT OR GRANT NUMBER(s) DAAL03-87-K-0003
9. PERFORMING ORGANIZATION NAME AND ADDRESS Texas A&M University Department of Statistics College Station, TX 77843-3143		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE May 1989 13. NUMBER OF PAGES 252
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

NA

18. SUPPLEMENTARY NOTES

The view, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

quantile data analysis, two sample nonparametric testing; comparison density function, components, boundary kernel density estimation, boundark kernel, goodness of fit statistics, empirical process. Theses, Ph.D.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)* The focus of this work is to derive functional and graphical statistical techniques for the two sample problem suitable for implementation in modern conputing environments. In the two sample problem, it is desired to test the null hypothesis that two independent random samples have a common distribution function. Assuming certain conditions on the distribution functions, a procedure is proposed which has strong graphical elements, a sound theoretical foundation, and estimates the relation of the two distributions if the null hypothesis is rejected. The proposed procedure has as its motivation the estimation of the comparison density and inference concerning its uniformity. keywords

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

ACKNOWLEDGEMENT

No dissertation is completed in a vacuum: it depends on the existing work in the area and on the training of the individual. Lively and interested discourse with one's colleagues is also an essential ingredient. In this respect, I am indebted to many of the faculty of the Department of Statistics, not only for discussions related directly to this work but for their time and interest while teaching the various courses I attended. In this regard, I would like to thank Tom Wehrly, Mike Longnecker and Ron Hocking from whom I have learned much. Thanks are also due Jeff Hart who has always made himself available for discussion. I would also like to thank Frits Ruymgaart who visited Texas A&M during the spring and summer of 1988. He taught a very interesting and useful class on empirical processes, and I benefited from his insights.

Graduate school can be a very good or bad experience depending on one's classmates. Scott Grimshaw, a fellow student and grandson of Loève, has helped it to be a very good experience. Special thanks are given to Joe Newton, a good friend as well as colleague, with whom I have spent many happy hours discussing computational statistics. Finally, I would like to express my greatest of thanks, appreciation, and respect to my advisor, Manny Parzen. It was he who inspired me to make the change to statistics from agricultural economics. It is not often that one has the opportunity to work with a man of such vitality and intelligence. If I have learned even a fraction of what one could learn from him, then success is all but assured.

This work would never have been undertaken were it not for the understanding, love, and steadfast support of my wife, Catheryn. She supported my move to statistics a mere fortnight after we were married and so confined herself to the life of a graduate student for an additional three and one half years. Yet she did so happily. In graduate school one learns things of a narrow and technical nature; Catheryn has shown me much of what is fundamentally true and important in this world. To her are due all my love and devotion.

ABSTRACT

Boundary Kernel Estimation

of the Two Sample Comparison Density Function. (May 1989)

William Pyle Alexander, B.S.A., University of Arkansas

Chair of Advisory Committee: Dr. Emanuel Parzen

The focus of this work is to derive functional and graphical statistical techniques for the two sample problem suitable for implementation in modern computing environments. In the two sample problem, it is desired to test the null hypothesis that two independent random samples have a common distribution function. Assuming certain conditions on the distribution functions, a procedure is proposed which has strong graphical elements, a sound theoretical foundation, and estimates the relation of the two distributions if the null hypothesis is rejected. The proposed procedure has as its motivation the estimation of the comparison density and inference concerning its uniformity.

The proposed procedure is both a statistical test of the null hypothesis and a model selection criterion. The test is based on components of a new stochastic process which is termed the kernel density process. This process is based on a boundary kernel estimate of the comparison density. It is proposed to apply a new test, the subset chi-square test, to these components. If the null hypothesis is rejected, the components found to be significant are used to construct a damped orthogonal series estimate of the comparison density.

The power of the proposed test under local alternatives is compared to two commonly used portmanteau statistics, the Cramér-von Mises and the Anderson-Darling, and to a third statistic suggested by this work. A new method for finding the power of these statistics under local alternatives is given. This method uses the fast Fourier transform to invert an approximation to the characteristic function of the statistic. The proposed test is seen to have good power properties. A simulation study is conducted to examine its small sample size. Its size is found to remain close to its nominal value.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

1. INTRODUCTION

1.1. The Two Sample Problem

The statistical analysis of two samples occupies a fundamental role in statistics. Data are collected under two regimes. Treatment and control, two time periods, two lots of goods, two levels of a concomitant variable, or two formulations of a product are just a few examples of such regimes. In each case, the researcher wishes to know whether the two data sets arise from the same underlying population. That is, are the two regimes the same? The number of statistical texts at all levels which treat the problem attests to the fundamental nature of the question: see, for example, Keller, Warrack and Bartel (1988) (undergraduate level methods), Montgomery (1984) (graduate level methods), Hocking (1985) (graduate level linear model theory), Randles and Wolfe (1979) (graduate level nonparametric theory), and Kendall and Stuart (1979) (graduate level parametric theory).

That the two sample problem has old and venerable roots can be seen from the writings of the great statistician Sir R.A. Fisher. Fisher (1948), page 122, contrasts the importance of testing a single mean versus the equality of two means in experimental work under the assumption of normality and equal variances: "in experimental work it is even more frequently necessary to test whether two samples differ significantly in their means, or whether they may be regarded as having arisen from the same population." Fisher's comments relate not only the relative importance of the two sample problem, but also something of its age. The statistic Fisher proceeds to discuss is the well known Student's t. Concerning the distribution of t, Fisher notes on page 16 that "it is equally fortunate that the distribution of t, first established by 'Student' in 1908, in his study of the probable error of the mean, should be applicable, not only to the case there treated, but to the more complex, but even more frequently needed, problem of the comparison of two mean values."

The format and style follows that of *The Annals of Statistics*.

While dealing with the two sample problem in various contexts over the years, statisticians have proposed an impressive number of tests. The *t*-test, Wilcoxon, median, normal scores, Kolmogorov-Smirnov, Anderson-Darling, and Cramér-von Mises are just a few of many. These tests range from making quite specific assumptions on the nature of the two distributions to almost none. The sorts of assumptions commonly made are discussed in Section 2.

Formally, the two sample problem can be stated in the following terms. Let $X = (X_1, \ldots, X_m)$ be a random sample from a population with distribution function $F(x) = P[X_i \leq x]$ for $i = 1, \ldots, m$. Let $Y = (Y_1, \ldots, Y_n)$ be an independent random sample from a population with distribution function $G(x) = P[Y_j \leq x]$ for $j = 1, \ldots, n$. Stating $X_1, \ldots, X_m$ is a random sample from F means that $X_1, \ldots, X_m$ are independent and identically distributed (iid) according to the distribution function F. These properties are nearly universally assumed. Some work has been done to relax the assumption of independence within each sample [see, for example, Harpaz (1985)]. However, in this work the usual assumption of independence will be made. The mathematical representation of the null hypothesis that the two samples arise from the same population is $H_o: F = G$. By this is meant that $F(x) = G(x)$ for $-\infty < x < \infty$.

Since there already exists such a plethora of two sample tests, it is only natural to ask why another is needed. The answer to this question is threefold. First, it is the goal of this research to derive a unified data exploratory method. That is, a methodology is sought which makes minimal assumptions on F and G *a priori* and will to the greatest extent possible let the data determine the outcome.

Second, as typically implemented, two sample techniques possess no graphical elements. They are statistical tests which return only an accept or reject response. The advent of the personal computer and workstations brings the potential to substantially alter the way in which statistical analysis is conducted. In particular, there is great potential for a more graphical, exploratory and flexible approach to data analysis. Packages such as *Timeslab* [Newton (1988)] for the IBM personal computer family and *S* [Becker, Chambers, and Wilks (1988)] for

UNIX workstations are appearing to realize this potential. These new computing environments open broad new areas of statistical research into methodologies to exploit them. This research shall develop tests and estimators for the two sample problem which are better suited to these environments than existing ones.

Third, the relevant standard two sample statistics give no indication of the actual relation of F to G should the null hypothesis be rejected. This shall be seen in Section 2. The most natural question to ask after the null hypothesis is rejected is "Why was it rejected?". Yet surprisingly, most techniques are silent on this point. Graphics enter here as the natural response. A graph should be an estimate of some sort of the relation of F to G. In this context, many practicing statisticians might plot the empirical distribution functions of the two samples. This and similar procedures are, at best, *ad hoc*. A particular statistic rejects H_o and one then proceeds to examine a picture of functions which are either step functions or piece-wise continuous to try to discern why. What is sought here is a unified approach where the graph and the test are derived from a common foundation.

In summary, the purpose of this research is to derive a new procedure for the two sample problem. This is to be a computer intensive, graphical and data-exploratory procedure which makes minimal assumptions on the character of the distribution functions, F and G. Further, as part of the framework, should the null hypothesis be rejected, it is required that some explanation be given. This explanation, in the form of a graph, should describe the relation of F and G.

1.2. Outline of This Dissertation

This dissertation is divided into six sections and two appendices. Section 2 is a review of the literature. Existing approaches to the two sample problem are examined first. Such approaches include parametric and nonparametric tests and tests against specific and general alternatives. Linear rank statistics and the Cramér-von Mises and Anderson-Darling statistics are examined in depth. The comparison distribution and density functions and related stochastic processes are defined and their properties enumerated. Existing techniques based on these

quantities are reviewed. It is seen that a nonparametric method is appropriate for estimating the comparison density. This leads to an in-depth review of such methods and the selection of the Gasser-Müller boundary kernel.

Section 3 examines the properties of the Gasser-Müller boundary kernel estimator of the comparison density. Conditions for its asymptotic normality under H_o and pointwise consistency under any alternative are derived when the bandwidth shrinks to zero. The kernel density process is defined and its weak convergence to a Gaussian process is proved for a fixed bandwidth. Components of this process are defined in terms of the inner product of the process and the eigenfunctions of its covariance kernel under H_o. These components are shown to be appropriate for testing the null hypothesis. They have interpretations both as generalized Fourier coefficients and as rank statistics. Their asymptotic distribution under H_o is derived. A new test, called the subset chi-square test, is applied to the components. This test, in turn, suggests an orthogonal series estimator of the comparison density based on the eigenfunctions. Finally, recommendations for the choice of bandwidth for the boundary kernel estimator are made.

Section 4 examines the power and size of the methods derived in Section 3. Conditions for the weak convergence of the kernel density process under local alternatives are established. Power functions for the Cramér-von Mises and Anderson-Darling statistics are found by using a fast Fourier transform (FFT) to numerically invert the characteristic function. Power functions for the subset chi-square test are found by simulation. The methods of Section 3 are seen to have very good power characteristics. Since the subset chi-square test is based on the asymptotic distribution of the components, a simulation is conducted to gauge the size of the test in small samples. The procedure is found to maintain a reasonable size even for small samples.

Section 5 applies the techniques of Section 3 to two data sets. One data set consists of observed data and the other of simulated data. Section 6 presents a summary and conclusions and outlines areas of future research. Appendix A is a glossary of notation and Appendix B gives proofs of the theorems stated in Sections 3 and 4.

2. REVIEW OF THE LITERATURE

2.1. Introduction

Section 2 is a review of the existing methodologies which this research touches or builds upon. Subsection 2.2 reviews the methods of two sample analysis as usually employed by the statistics community. The subsection closes with a statement as to the criteria any methodology derived in this work must meet and the rationale for these criteria.

Subsection 2.3 reviews a concept known as the comparison density. It is argued that a methodology based on the comparison density fulfills the criteria outlined in Subsection 2.2. The stochastic processes upon which any technique must be based are reviewed and links between them refined. Existing techniques of estimating the comparison density are reviewed. Each of these will be compared to the criteria outlined in Subsection 2.2. They will be seen to fall short of fulfilling all the criteria outlined, but will prove to be valuable stepping stones in this work.

Subsection 2.4 reviews nonparametric density estimation techniques for densities having support [0,1]. Such a technique will be employed to estimate the comparison density in this work. Density estimates on compact support pose special problems which will be discussed in detail. At the end of the subsection, an estimation technique is selected.

2.2. Review of Two Sample Techniques

2.2.1. *Introduction.* Given a random sample, $X_1, \ldots, X_m$, from a continuous distribution function, F, and an independent random sample, $Y_1, \ldots, Y_n$, from a continuous distribution function, G, it is desired to test the null hypothesis $H_o: F = G$. In this subsection, existing tests of this hypothesis are reviewed. There are many tests that have been suggested and used over the years. Some are applicable to very special and specific distributional assumptions concerning the two samples. Others are applicable to very general cases. Yet, as implemented,

none lend themselves fully to the type of computer intensive, graphical, functional and data exploratory approach outlined in Section 1.

There are two schemes by which two sample tests can be classified which will be discussed here. The first method classifies a test as to whether it is parametric or nonparametric. The second classifies a test as to whether the alternate hypothesis is general or specific. These are discussed in Subsections 2.2.2 and 2.2.3, respectively. Subsection 2.2.4 reviews rank statistics, which are nonparametric tests against specific alternatives. Subsection 2.2.5 reviews nonparametric tests against general alternatives. Having reviewed the commonly used methodologies, a list of criteria for the ideal method is proposed in Subsection 2.2.6. This list summarizes the desirable properties a methodology should have to more fully utilize computer and graphic intensive modes for data analysis.

2.2.2. *Parametric and Nonparametric Tests.* Tests can be classified as to whether they are parametric or nonparametric. In this subsection, parametric and nonparametric tests are compared. It is seen that parametric tests are too restrictive for the goals of this research. Any tests derived will be nonparametric in nature.

Parametric tests assume that both F and G belong to a family of distributions, $\mathcal{F}_\theta$, which is indexed by a parameter θ. That is, one can write

$$\mathcal{F}_\theta = \{F(x) = F(x;\theta_1); G(x) = G(x;\theta_2) : \theta_1, \theta_2 \in \Theta \subseteq \mathbb{R}^k\}.$$

It is usually assumed that $\mathcal{F}_\theta$ is indexed in such a manner that for $F(x;\theta_1)$ and $G(x;\theta_2)$ in $\mathcal{F}_\theta$, one has $F(x;\theta_1) = G(x;\theta_2)$ for all x if and only if $\theta_1 = \theta_2$. This uniqueness property permits the reduction of the general two sample hypothesis of H_o: $F = G$ to H_o: $\theta_1 = \theta_2$.

With the two sample hypothesis reduced to testing the equality of two vector valued parameters, standard techniques from parametric inference may be brought to bear upon the problem. Tests such as the likelihood ratio, efficient score, Wald, and uniformly most powerful unbiased may potentially be derived. See Silvey (1975), Kendall and Stuart (1979) or Bickel and Doksum (1977) for background on these.

As an example of a parametric test, consider the two sample t-test. Let $\theta = (\mu, \sigma^2)$ and $\Phi(x)$ be the standard normal distribution function. Define the parametric family as,

$$\mathcal{F}_\theta = \{F(x) = \Phi((x-\mu_1)/\sigma); G(x) = \Phi((x-\mu_2)/\sigma) : \mu_1, \mu_2 \in \mathbb{R}; \sigma^2 > 0\}.$$

The null hypothesis is H_o: $\mu_1 = \mu_2$. A likelihood ratio test rejects H_o if

$$\lambda = \sup_{\mu_1=\mu_2 \in \mathbf{R}, \sigma^2>0} L(\mu_1, \mu_2, \sigma^2) / \sup_{\mu_1,\mu_2 \in \mathbf{R}, \sigma^2>0} L(\mu_1, \mu_2, \sigma^2)$$

is too small, where $L(\mu_1, \mu_2, \sigma^2)$ is the joint likelihood function of (μ_1, μ_2, σ) given $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ and is given by

$$L(\mu_1, \mu_2, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{(n+m)/2}} \exp\Big[-\frac{1}{2\sigma^2}\Big(\sum_{j=1}^{m}(x_i - \mu_1)^2 + \sum_{j=1}^{n}(y_i - \mu_2)^2\Big)\Big].$$

The likelihood function gives an instantaneous measure of the probability content at the point (μ_1, μ_2, σ^2) given the data $x_1, \ldots, x_m$, $y_1, \ldots, y_n$. In the case at hand, it is not hard to show that rejecting H_o for small values of λ is equivalent to rejecting H_o for large values of

$$\left| \frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) S_p^2}} \right|,$$

where $S_p^2 = \frac{1}{n+m-2}(\sum_{j=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_i - \bar{Y})^2)$ is the sample pooled variance and $\bar{X}$ and $\bar{Y}$ are the means of the first and second samples, respectively. This statistic is the standard two sample t-statistic.

Most parametric tests are generated in an analogous manner. One starts by writing down the likelihood function. The likelihood ratio, Wald and efficient scores tests are constructed around the maximum likelihood estimates of the parameters. Uniformly most powerful unbiased tests can be obtained in special cases and require an appeal to certain theorems detailing their existence and construction.

Nonparametric tests assume that F and G lie in a class, $\mathcal{F}$, which is so broad that it cannot be indexed by a finite dimensional parameter. An example

of such a class is the class of all continuous distribution functions. Nonparametric techniques rely heavily on transformations of the original random variables, $X_1, \ldots, X_m, Y_1, \ldots, Y_n$, to new random variables, $R_1, \ldots, R_m, S_1, \ldots, S_n$, such that when the null hypothesis is true these new random variables always have the same, known distribution. A set of random variables possessing this property is said to be nonparametric distribution-free under the null hypothesis. Tests of H_o are then based on $R_1, \ldots, R_m, S_1, \ldots, S_n$ rather than on $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. An example of such a transformation is the rank transformation. For the rank transformation, R_i is the rank of X_i and S_i is the rank of Y_i in the pooled or combined sample. When one pools the sample, $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are treated as being from one random sample, $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ of size $N = m + n$. The rank, R_i, is given by

$$R_i = \sum_{j=1}^{m} I(X_i \leq X_j) + \sum_{j=1}^{n} I(X_i \leq Y_j),$$

where $I(\cdot)$ is an indicator function which is 1 if the condition in parentheses is true and zero otherwise. There should be no ties in the ranks since it is assumed that F and G are continuous. Under the null hypothesis, $(R_1, \ldots, R_m, S_1, \ldots, S_n)$ are uniformly distributed over all $N!$ permutations of $1, \ldots, N$ and R_i is marginally distributed as uniform over $1, \ldots, N$ [see Lehmann (1975), page 58, for example].

The rank transform and tests based upon it are examined in detail in Subsection 2.2.4. For now, consider the Wilcoxon statistic which is given by

$$W = \sum_{j=1}^{m} R_i.$$

As the ranks are uniformly distributed over the $N!$ permutations under H_o, it can be shown [see Randles and Wolfe (1979), page 45] that $E[W] = m(N+1)/2$ and $Var[W] = nm(N+1)/12$. The full distribution of W under H_o can be found by enumeration or by asymptotic approximation.

There are three issues to be addressed in any discussion of parametric versus nonparametric tests. These are specification error, size, and power. A specification error occurs whenever F or G does not fall in the assumed parametric class,

$\mathcal{F}_\theta$. The size of a test is the probability of rejecting the null hypothesis when it is true. The power of a test is the probability of rejecting the null hypothesis when it is false. The tradition in statistics since the time of Fisher and Neyman has been to fix the size at some small value. For a given size, one then prefers a test which is more powerful. In most situations, there won't exist a test which is uniformly more powerful (UMP) than any other test. There usually aren't even what are called uniformly most powerful unbiased tests (UMPU). Such a test, when it exists, is most powerful in the class of unbiased tests. A test is unbiased if it has power greater than its size for all alternatives.

In the current context, these issues will be discussed via Table 1. Table 1 gives the size and power of the two sample t-test and the Wilcoxon test for 4 choices of m and n and 5 choices of $(\mu_1 - \mu_2)/\sigma$ and for $\mathcal{F}_\theta$ equal to the normal distribution family and to the Cauchy distribution family. This table is a partial reproduction of a table found on page 118 of Randles and Wolfe (1979). The size of the test falls under the column value of 0 for $(\mu_1 - \mu_2)/\sigma$. Each test is conducted to have nominal size 0.05. The power of the test is given under the remaining columns, for each choice of family, in increasing values of $(\mu_1 - \mu_2)/\sigma$. This table was created by simulation methods; see Randles and Wolfe for further details on its construction.

If a parametric assumption is valid, one expects the appropriate parametric test to be at least as powerful as a nonparametric test. This is so simply because one is bringing more information to bear on the problem. The nonparametric test must protect against a huge array of possible underlying F and G which the parametric test ignores. This is borne out by Table 1 where the t-test is seen to be more powerful than the Wilcoxon for the normal family. However, it is also important to notice that this difference is slight.

The t-test will experience specification error when the Cauchy family holds whereas the Wilcoxon will not. The implications of the t-test experiencing a specification error are demonstrated under the Cauchy heading of Table 1. The size of the t-test shows large fluctuations away from 0.05. The Wilcoxon test shows no such failing. Finally, notice that the Wilcoxon test is now much more

Table 1

Empirical power and size of the t-test and the Wilcoxon linear rank test for the normal and Cauchy families. T is the t-test and W is the Wilcoxon test.

			$(\mu_1 - \mu_2)/\sigma$				
m	n	Test	0	0.3	0.6	0.9	1.2
			Normal				
5	5	T	0.044	0.111	0.213	0.356	0.523
		W	0.046	0.108	0.208	0.346	0.503
15	15	T	0.052	0.206	0.497	0.785	0.947
		W	0.054	0.205	0.479	0.766	0.933
5	15	T	0.047	0.144	0.303	0.511	0.724
		W	0.048	0.141	0.287	0.492	0.694
15	5	T	0.053	0.149	0.313	0.518	0.729
		W	0.050	0.140	0.296	0.499	0.703
			Cauchy				
5	5	T	0.024	0.066	0.132	0.207	0.288
		W	0.051	0.118	0.218	0.323	0.408
15	15	T	0.030	0.079	0.153	0.243	0.333
		W	0.046	0.210	0.484	0.700	0.839
5	15	T	0.056	0.087	0.137	0.205	0.282
		W	0.046	0.133	0.284	0.441	0.576
15	5	T	0.061	0.097	0.146	0.209	0.279
		W	0.046	0.140	0.297	0.457	0.590

powerful than the *t*-test. The superiority of the Wilcoxon over the *t*-test is much greater in this case than that of the *t*-test over the Wilcoxon when normality holds.

In summary, one can surmise that in finite samples, parametric tests will often outperform nonparametric tests in situations where the parametric test is appropriate. In such circumstances, should one have just cause to assume a parametric model one should surely do so. However, parametric tests are sensitive to specification errors which nonparametric tests do not experience. As it is the goal of this work to develop a procedure which applies to as broad an underlying class of distributions as possible, parametric tests will not be

considered. However, the methods derived will be competitive with parametric tests. Nonparametric tests are discussed at greater length in Subsections 2.2.4 and 2.2.5.

2.2.3. *General and Specific Alternative Hypotheses.* In this subsection, types of alternative hypotheses are examined. The alternative hypothesis specifies the set of possible relations of F to G should the null hypothesis not be true. Although ignored until now, the alternative hypothesis must always be stated. The class of alternatives considered will have a profound effect upon the properties of a test. It is seen in this subsection that for the type of procedure to be constructed, a general and not specific alternative is needed.

For two sample tests, the alternative hypothesis can be divided into roughly two categories. The first is a general alternative and the second a specific alternative. A general alternative leaves the fashion in which F and G are related unspecified. A specific alternative will place some structure on the manner in which F and G are allowed to differ. The type of alternative considered relates back to the underlying class, $\mathcal{F}$, to which F and G belong. A few examples should shed some light on this. Consider any parametric test so that $F(x;\theta_1)$, $G(x;\theta_2) \in \mathcal{F}_\theta$. The null hypothesis of $H_o: \theta_1 = \theta_2$ is complemented by an alternative of the form $H_a: \theta_1 \neq \theta_2$ or $H_a: \theta_1 > \theta_2$. The important point is that F and G are still in $\mathcal{F}_\theta$ even under the alternative. The alternative is said to be specific.

Most nonparametric tests are constructed against specific alternatives, also. The most common alternative is the location alternative. For the location alternative, the class of distributions is defined by

$$\mathcal{F} = \{F(x) = H(x); G(x) = H(x-\theta) : H \text{ is a continuous d.f., } \theta \in \mathbb{R}\}.$$

The null hypothesis reduces to $H_o: \theta = 0$, yet the test must still be nonparametric because H is any continuous distribution function, a class too broad to be indexed. The alternative hypothesis in this setting can be $H_a: \theta \neq 0$ or $H_a: \theta > 0$. The alternative is specific because F and G are related through H by θ. A third and final example of a specific alternative is the scale alternative. The family for

scale alternatives is

$$\mathcal{F} = \{F(x) = H(x); G(x) = H(x/\eta) : H \text{ is a continuous d.f., } \eta > 0\}.$$

Again, the class is still too broad to be indexed but does place structure on the relationship of F to G under the alternative. Nonparametric tests of location include the Wilcoxon, median and normal scores (location) rank tests. Nonparametric tests of scale include the Mood and normal scores (scale) rank tests.

A general alternative leaves the way in which F and G are related unspecified. A typical class in this case might be

$$\mathcal{F} = \{F \text{ is a continuous d.f.}; G \text{ is a continuous d.f.}\}.$$

One can see that this is a much broader class than the location and scale alternatives considered above. Minimal assumptions are made on the true relation of F to G under H_a.

The class of alternatives is important for just the same reasons as discussed in Subsection 2.2.3 on the relation of parametric to nonparametric tests. The issues are power and specification. A test against a specific alternative will usually be more powerful than a test against a general alternative if the alternative which actually holds falls in the class considered by the former. On the other hand, a specific alternative can fail miserably if the true alternative falls outside the class for which it is designed. For example, using the techniques discussed in Section 4, it can be shown that asymptotically the Wilcoxon test has power equal to its size if a local scale alternative holds and the underlying distribution is symmetric about zero.

As with parametric tests, if one has justification to use a test designed against a specific alternative, it should by all means be used. However, since the purpose of this work is to design a methodology which makes minimal assumptions on F and G, broader alternatives than location or scale must be considered. Tests against such broader alternatives are called omnibus or portmanteau.

2.2.4. *Nonparametric Tests Against Specific Alternatives (Linear Rank Statistics).* In this subsection, linear rank statistics, which are nonparametric

tests against specific alternatives, are reviewed. It has been decided in Subsection 2.2.3 not to employ tests against specific alternatives such as rank tests. However, rank tests will be seen to have a role to play and as such merit discussion. Rank statistics have many equivalent or asymptotically equivalent representations. In this subsection, the notation of Chernoff and Savage (1958) is used. They define a rank statistic having the form

$$
\begin{aligned}
S_N &= \int J_N[H_N(x)]dF_m(x) \\
&= \frac{1}{m}\sum_{j=1}^{m} J_N\Big(R_i/N\Big),
\end{aligned}
\tag{2.2.1}
$$

where R_i is the rank of X_i in the pooled sample, J_N is known as a score function, unqualified integrals are assumed to be taken over the real line, F_m is the sample or empirical distribution function of the first sample,

$$F_m(x) = \frac{1}{m}\sum_{j=1}^{m} I(X_i \le x)$$

and H_N is the sample distribution function of the pooled sample,

$$
\begin{aligned}
H_N(x) &= \frac{1}{N}\Big(\sum_{j=1}^{m} I(X_i \le x) + \sum_{j=1}^{n} I(Y_i \le x)\Big) \\
&= \lambda_{(N)}F_m(x) + (1-\lambda_{(N)})G_n(x).
\end{aligned}
$$

The sample distribution function of the second sample is $G_n(x)$ and $\lambda_{(N)} = m/(m+n) = m/N$ is the fraction of the pooled sample represented by the first sample. The sample distribution function of the pooled sample estimates the population quantity $H(x) = \lambda_{(N)}F(x) + (1-\lambda_{(N)})G(x)$.

As might be expected, the small sample distribution of S_N under H_o or H_a is not always easy to determine. This is true even of the Wilcoxon statistic. The finding of percentage points of the distribution of S_N is greatly simplified by the celebrated work of Chernoff and Savage (1958). Using what has come to be called a Chernoff-Savage approach, they demonstrate the asymptotic normality of S_N. For simplicity, assume that $J_N(u) = J(u)$ does not depend on N. One then has

Theorem 2.2.1 (Chernoff and Savage, 1958). *If $J(u)$ is not constant and if $|J^{(i)}(u)| \leq K|u(1-u)|^{-i-(1/2)+\delta}$ for $i = 0, 1, 2$ and some K and $\delta > 0$, then for fixed and continuous F and G, one has S_N is AN(μ,σ_N^2), where*

$$\mu = \int J[H(x)]dF(x)$$

and

$$N\sigma_N^2 = 2(1-\lambda_{(N)})\Big\{\int\int_{x<y} G(x)[1-G(y)]J'[H(x)]J'[H(y)]dF(x)dF(y) + \frac{1-\lambda_{(N)}}{\lambda_{(N)}}\int\int_{x<y} F(x)[1-F(y)]J'[H(x)]J'[H(y)]dG(x)dG(y)\Big\}$$

providing $\sigma_N \neq 0$.

The notation S_N is AN(μ,σ_N^2) means that the distribution function of the random variable $(S_N - \mu)/\sigma_N$ converges pointwise to the distribution function of a standard normal random variable. To find approximate values of the distribution function of S_N, one need only calculate the values of μ and σ_N. In many practical circumstances, the values of μ and σ_N^2 can be worked out. For example, taking $J(u) = u$ (Wilcoxon scores), under H_o one finds

$$\mu = \int F(x)dF(x) = \int_0^1 u du = 1/2$$

and

$$\begin{aligned} N\sigma_N^2 &= 2\frac{1-\lambda_{(N)}}{\lambda_{(N)}}\int\int_{x<y} F(x)[1-F(y)]dF(x)dF(y) \\ &= 2\frac{1-\lambda_{(N)}}{\lambda_{(N)}}\int [1-F(y)]dF(y)\int_{-\infty}^{y} F(x)dF(x) \\ &= \frac{1-\lambda_{(N)}}{\lambda_{(N)}}\int F(y)^2[1-F(y)]dF(y) \\ &= \frac{1}{12}\frac{1-\lambda_{(N)}}{\lambda_{(N)}}. \end{aligned}$$

By the Chernoff-Savage theorem, one can conclude that the Wilcoxon statistic is AN$(1/2, (1-\lambda_{(N)})/(12N\lambda_{(N)}))$.

Table 2

Commonly used score functions for linear rank statistics, location and scale alternatives.

Name	Score Function
Location	
Wilcoxon	u
Normal	$\Phi^{-1}(u)$
Median	$I(u < 1/2)$
Scale	
Mood	$(u - 1/2)^2$
Normal	$\Phi^{-1}(u)^2$
Ansari-Bradley	$\|u - 1/2\|$

Many different forms of the score function have been proposed, some which depend on N and some which do not. It is the score function which determines the properties of the statistic. It is up to the user of these techniques to choose the score function. Some considerations for its choice are now given. Table 2 gives some commonly used score functions for scale and location alternatives. Notice that the score functions corresponding to location alternatives are monotone and those corresponding to scale alternatives have one sign change in their derivative. If one were to redefine the score function as $J(u) = J(u) - \mu$ (thus centering the statistic), this observation can be recast in terms of zero crossings. The score functions for location alternatives have one zero crossing in $(0, 1)$ and those for scale alternatives have two. Eubank, LaRiccia and Rosenstein (1987) give further intuition into this matter. For now, it is enough to notice such a pattern.

Given the relative freedom in the choice of score function, one might ask if it is possible to choose it optimally in some fashion. The answer is yes in the following sense. For location alternatives, $F(x) = H(x)$ and $G(x) = H(x - \theta)$, the optimal score function is

$$J_L(u) = -\frac{h'Q^H(u)}{hQ^H(u)},$$

where $h = H'$ is the density function of the distribution function H; $Q^H(u) = \inf\{x: F(x) \geq u\}$ is the quantile function of H and $hQ^H(u) = h[Q^H(u)]$ is a

composite function as is $h'Q^H(u)$. The density, h, and its derivative, h', are assumed to exist. The score function $J_L(u)$ is optimal in the sense that it maximizes the asymptotic relative efficiency (ARE) of the test as defined by Noether [see Randles and Wolfe (1979) pages 147 ff.]. The asymptotic relative efficiency gives a measure of the power of one test relative to another. In fact, the test that results from using J_L is asymptotically relatively efficient. This means that no test, not even a parametric one, will produce a better ARE. The optimality of J_L is quite strong. Similarly, for scale alternatives one finds that the optimal score function is

$$J_S(u) = -Q^H(u)\frac{h'Q^H(u)}{hQ^H(u)}.$$

Applying these formulas, one sees that the Wilcoxon is optimal at detecting location shifts if H is the logistic distribution and the normal scores tests are optimal at detecting location and scale shifts in underlying normal populations. In one sense it is a drawback that to achieve the optimality one must know the underlying family of distributions, H, to which F and G belong. This is not the case if one is merely interested in protecting best against certain classes of distributions. For example, if one thought that the underlying distribution might be slightly longer tailed than the normal, the Wilcoxon test is a good choice. Even though it is optimal only for H logistic (which has slightly longer tails than the normal), one should expect it to perform well against the broader family. Further, since the test is nonparametric one is protected in case F and G are strongly non-logistically distributed.

2.2.5. *Nonparametric Tests Against General Alternatives.* As has been shown in Subsections 2.2.2 and 2.2.3, the class of nonparametric tests against general alternatives most nearly matches the goal of minimal assumptions about F and G outlined in Section 1. In this subsection, existing nonparametric tests against general alternatives are reviewed. In addition to standard properties, it should be examined how these tests fit into a computer oriented data exploratory environment. It is seen that as simple statistics, they do not fit well into such an environment. Two tests are examined in detail: the Cramér-von Mises and the

Anderson-Darling. The Kolmogorov-Smirnov test is briefly mentioned, but not examined in detail.

The Cramér-von Mises test has been defined in a number of ways, none quite the same but all having the same spirit. Lehmann (1951) and Rosenblatt (1952) define the Cramér-von Mises statistic as

$$\frac{nm}{2N}\int [F_m(x) - G_n(x)]^2 d[F_m(x) + G_n(x)];$$

Kiefer (1959) and Fisz (1960) employ the definition

$$\frac{nm}{N}\int [F_m(x) - G_n(x)]^2 dH_N(x);$$

the Pyke and Shorack (1968) process leads to the definition

$$N\frac{\lambda_{(N)}}{1-\lambda_{(N)}}\int_0^1 [F_m Q_N^H(u) - u]^2 du,$$

where $Q_N^H(u)$ is the sample quantile function of the pooled sample; and Parzen's (1983) definition of the comparison distribution function leads to the definition

$$\text{CVM}_N = N\frac{\lambda_{(N)}}{1-\lambda_{(N)}}\int_0^1 [D_N(w) - w]^2 dw, \tag{2.2.2}$$

where $D_N(w)$ is the sample distribution function of the normalized ranks, $R_1/N, \ldots, R_m/N$. This last definition is the one which is used throughout this work. All of these versions have the same motivation: one is measuring the distance of F to G by an integral of a squared function. Here only (2.2.2) is considered in detail.

The limiting distribution of (2.2.2) is the same as that of the corresponding one sample statistic and is given by Anderson and Darling (1952). The limiting distribution depends on that of the integrand which is viewed as a continuous parameter stochastic process. To achieve a limiting process for the integrand, one must make certain additional assumptions on F and G. These are detailed in Subsection 2.3.3 and won't be discussed here further.

Durbin and Knott (1972) give a very important alternate representation for the Cramér-von Mises statistic in terms of what they call components. Although

they use the one sample problem as a format, their procedures apply to the two sample problem as well. Using their techniques, one finds that

$$\text{(2.2.3)} \qquad \text{CVM}_N = \sum_{j=1}^{\infty} \frac{1}{j^2\pi^2} Z_{Nj}^2,$$

where the

$$Z_{Nj} = j\pi\sqrt{2}\int_0^1 \sin(\pi j w)\sqrt{\frac{N\lambda_{(N)}}{1-\lambda_{(N)}}}[D_N(w) - w]dw$$

are referred to as the components of the Cramér-von Mises statistic. The components, $Z_{N1}, Z_{N2}, \ldots$, are asymptotically independent with the standard normal distribution under H_o.

Durbin and Knott's (1972) result is derived by an orthonormal expansion of the random function $\sqrt{N\lambda_{(N)}/(1-\lambda_{(N)})}[D_N(w) - w]$. The techniques used are basic to Fourier analysis. The equality (2.2.3) follows easily from Parseval's identity. The distributional results are somewhat deeper in nature and a good discussion is given in Shorack and Wellner (1986), pages 215 ff.

Anderson and Darling (1952) suggest an alternative statistic, AD_N, which can be written as

$$\text{AD}_N = \frac{N\lambda_{(N)}}{1-\lambda_{(N)}}\int_0^1 [D_N(w) - w]^2/w(1-w)dw.$$

The rationale for the extra term, $w(1-w)$, is to give each point of the process $\sqrt{N\lambda_{(N)}/(1-\lambda_{(N)})}[D_N(w) - w]$ equal weight in a statistical sense. This follows from the fact that under H_o its limiting process, call it $\text{CD}_N(w)$, has variance $w(1-w)$. Anderson and Darling (1952) determine the distribution of AD_N under the null hypothesis. Durbin and Knott show that this statistic, too, has a representation in terms of components. This representation is

$$\text{AD}_N = \sum_{j=1}^{\infty} \frac{1}{j(j+1)} Z_{Nj}^2,$$

where

$$Z_{Nj} = \int_0^1 \text{LP}_j(w)\sqrt{\frac{N\lambda_{(N)}}{1-\lambda_{(N)}}}[D_N(w) - w]/\sqrt{w(1-w)}dw,$$

with

$$LP_j(w) = 2\sqrt{(2j+1)w(1-w)}L_j'(2w-1),$$

and L_j is the j^{th} Legendre polynomial. Again, under H_o, the components are asymptotically distributed as independent standard normal random variables.

The asymptotic representations for CVM_N and AD_N allow important interpretations of the manner in which these statistics operate. It is well known that these tests are consistent against any alternative [see Randles and Wolfe (1979), page 384], yet one must certainly be sensitive to the issue of power as well. Consider the first component, Z_{N1}, of the Cramér-von Mises statistic. It can be written as

$$\begin{aligned} Z_{N1} &= a_N \int_0^1 \sin(\pi w)[D_N(w) - w]dw \\ &= b_N \int_0^1 \cos(\pi w)d[D_N(w) - w] \\ &= b_N \sum_{j=1}^{m} \cos(\pi R_j/N), \end{aligned}$$

where a_N and b_N are constants depending only on N. This component has the form of a linear rank statistic with score function $J_1(u) = \cos(\pi u)$. Since $J_1(u)$ has but one zero crossing in $(0,1)$, the first component is a test against a location shift. It can be shown that $J_1(u)$ is the optimal score function for detecting shifts in the Cauchy family. In the same manner, the first component of the Anderson-Darling statistic can be shown to be the Wilcoxon statistic. Similarly, the second component of CVM_N is a rank statistic with score function $J_2(u) = \cos(2\pi u)$ which is a score for scale alternatives. The process continues with score functions of successively higher frequency.

The interpretation of the components as rank statistics is very important. Both CVM_N and AD_N successively and rapidly downweight these rank statistics in calculating an overall portmanteau statistic. Although they may be consistent against general alternatives, one expects that they would have poor power characteristics against any but the first few components. This is indeed the case, as is seen in Section 4.

There does exist a third statistic, the Kolmogorov-Smirnov statistic, which is consistent against general alternatives. It is defined as

$$\mathrm{KS}_N = \sqrt{\frac{N\lambda_{(N)}}{1-\lambda_{(N)}}} \sup_{0\le w\le 1} |D_N(w) - w|,$$

which measures the deviations from uniformity in a supremum norm sense. It tests H_o only by the largest deviation of $D_N(w)$ from w. The Kolmogorov-Smirnov test does not have a representation in terms of components, which makes its response to various alternatives much more difficult to gauge.

As a final point concerning nonparametric tests against general alternatives, consider how these statistics relate to the criteria outlined in Section 1. As nonparametric tests against general alternatives, they certainly make minimal assumptions on F and G. There are certain troubling questions about their power. It is also difficult to see how they fit into a graphical, exploratory mode of data analysis. As simple statistical tests, they simply accept or reject. There is no explanation as to why H_o is rejected should it be.

2.2.6. *Criteria for a Methodology.* Having reviewed existing two sample techniques and armed with an outline of goals from Section 1, a list of criteria for a methodology can now be given. The list gives the desirable properties a procedure should possess in order to attain the goals given in Section 1. These criteria result directly from a combination of these goals and observations made concerning existing techniques in this section.

The criteria for a two sample procedure which must be met by any derived in this work are:

1. It is not solely number oriented but does possess graphical features.
2. It is not only a statistical test but also a selection procedure for a model of the relation of F to G.
3. It should be omnibus.
4. It should be nonparametric distribution free under the null hypothesis.

A procedure with strong graphical elements is desired to take advantage of modern computing environments. Statistics are needed, to be sure. However,

numbers by themselves cannot convey the quantity or diversity of information of which graphs are capable. Graphs have shape; numbers do not. Statistics are useful as diagnostics and indicators; however, there is no longer any need to rely upon them exclusively.

Dual to statistical testing is model selection. Suppose a model, M, for a process is to be chosen from the class of possible models, $\mathcal{M}$. The model gives, in some fashion, the true behavior of the process. In the two sample case, it would give the true relation of F to G. Suppose further that the null hypothesis corresponds to some subset, $\mathcal{A}$, of $\mathcal{M}$. Choosing a model is dual to testing the null hypothesis in that the null hypothesis is rejected if and only if the chosen model is not in $\mathcal{A}$. Similarly, a test of H_o can be viewed as a model selection process if by rejecting H_o an element of $\mathcal{M}$ not in $\mathcal{A}$ is selected. Any procedure derived must explicitly represent this duality.

As stated in Section 1, it is desired to make minimal assumptions on F and G either under H_o or H_a. In term of statistical terminology, it has been seen in this section that this desire translates into a nonparametric test against general alternatives. The wish to have a test consistent against any alternative is tempered by the desire for a test which has good power characteristics against a wide range of alternatives. It will not be required that a test be consistent against any alternative, but that it be consistent for a wide range of alternatives or be omnibus. The procedure should also be nonparametric distribution free under H_o so that the distributional problems possess a solution.

2.3. Review of the Comparison Density

2.3.1. *Introduction.* There is an object which lends itself to the sort of graphical, functional type of portmanteau test which was outlined in Subsection 2.2.6. This object is termed the comparison density by Parzen (1983). The estimation of and tests based on the comparison density will form the foundation of this dissertation. In the following subsections, the comparison density is defined, the properties of related stochastic processes are discussed, and existing tests based on it are reviewed.

2.3.2. *Definition of the Comparison Density.* In this subsection, the comparison density is defined and its properties reviewed. Let a_f and a_g be the lower endpoints of f and g, respectively, so that $a_f, a_g \geq -\infty$. Similarly, let b_f and b_g be the upper endpoints of f and g, respectively, so that $b_f, b_g \leq \infty$. From the outset, assume that the distribution functions, F and G, the quantile functions, Q^F and Q^G, and the densities, f and g, satisfy

(2.3.1)
a. Q^F and Q^G are continuous;
b. F and G are absolutely continuous with densities f and g;
c. f and g are both continuous on the interval $(a_f \wedge a_g, b_f \vee b_g)$.

Parzen (1983) defines the comparison distribution function to be,

$$D_\lambda(w) = FQ_\lambda^H(w) \equiv F \circ Q_\lambda^H(w) \text{ for } 0 \leq w \leq 1,$$

where $Q_\lambda^H(w)$ is the quantile function of $H_\lambda(x) = \lambda F(x) + (1-\lambda)G(x)$ and $\circ$ means function composition. A few of the simple properties of D_λ are: (a) $D_\lambda(0) = 0$; (b) $D_\lambda(1) = 1$; (c) D_λ is increasing on $[0,1]$; and (d) D_λ is absolutely continuous on $[0,1]$. These properties justify the term 'distribution' as D_λ is, in fact, a distribution function.

The comparison density is just the derivative of the comparison distribution function

$$d_\lambda(w) = \frac{d}{dw} D_\lambda(w) = \frac{fQ_\lambda^H(w)}{hQ_\lambda^H(w)}, \tag{2.3.2}$$

since $\frac{d}{dw} Q_\lambda^H(w) = 1/hQ_\lambda^H(w)$ [see Parzen (1979)]. Note that condition (2.3.1c) ensures that $d_\lambda(w)$ is continuous on $(0,1)$. The continuity of $d_\lambda(w)$ is needed to show the weak convergence of the comparison distribution process (see Subsection 2.3.3). The condition (2.3.1c) allows for many possible F and G, but some choices are excluded. For example, taking F as the N(0,1) distribution and G as the standard lognormal does satisfy this condition since G is continuous on $\mathbb{R}$. Taking F as the N(0,1) distribution and G as the standard exponential distribution does not satisfy condition (2.3.1c) since G is discontinuous at 0. The comparison density will also be discontinuous. Parzen (1983) gives several of the

elementary properties of d_λ as, (a) $0 \leq d_\lambda(w) \leq 1/\lambda$; (b) $d_\lambda(w) \to 0$ if $f \to 0$; and (c) $d_\lambda(w) \to 1/\lambda$ if $g \to 0$.

The most important interpretation of d_λ is that of a likelihood ratio. The comparison density is the likelihood ratio of the density of the first sample to the density of the pooled sample, f to h_λ, evaluated at the quantile function of the pooled sample, Q_λ^H. Now if $F = G$, $h_\lambda(x) = f(x)$ for all x. If $F \neq G$, then f will differ from g on at least an interval (since both are continuous). Consequently, $F = G$ if and only if $d_\lambda(w) = 1$ for $0 \leq w \leq 1$. Thus the hypothesis $H_o: F = G$ is equivalent to

$$H_o: d_\lambda(w) = 1 \text{ for } 0 \leq w \leq 1.$$

Furthermore, if the alternative $F \neq G$ holds then d_λ specifies the way in which the hypothesis fails. This specification is given by departures of d_λ from uniformity. It is also possible to specify these departures in terms of the usual likelihood ratio of f to g by noting, as Parzen (1983) does, that

$$\frac{1}{d_\lambda(w)} = \lambda + (1-\lambda)\frac{gQ_\lambda^H(w)}{fQ_\lambda^H(w)}.$$

However, there is really no need to go to this trouble unless an estimate of f/g is specifically desired. Visually, it is enough to know that $d_\lambda(w) > 1$ if and only if $fQ_\lambda^H(w) > gQ_\lambda^H(w)$. A further argument for interest in d_λ instead of f/g is that d_λ is bounded between 0 and $1/\lambda$ whereas f/g will often be unbounded. The estimation of unbounded functions is a significantly more difficult task which is best avoided, if possible.

Given a plot of d_λ, one might wonder if one can determine the kinds of F and G that generated it. Figure 1 presents d_λ for a variety of F and G. Figures (a), (c), (d) and (f) are location alternatives, that is, $G(x) = F(x-\theta)$, for some constant θ. Figure (b) is a scale alternative, so $G(x) = F(x/\theta)$. Although one might be tempted to classify Figure (e) as a scale alternative based on the fact that $G(x) = F(x/2)$, it is best characterized as a location alternative. This is so since it is easily converted to a location alternative by taking logarithms of the random variables. This interpretation is doubly pleasing since Figure (e)

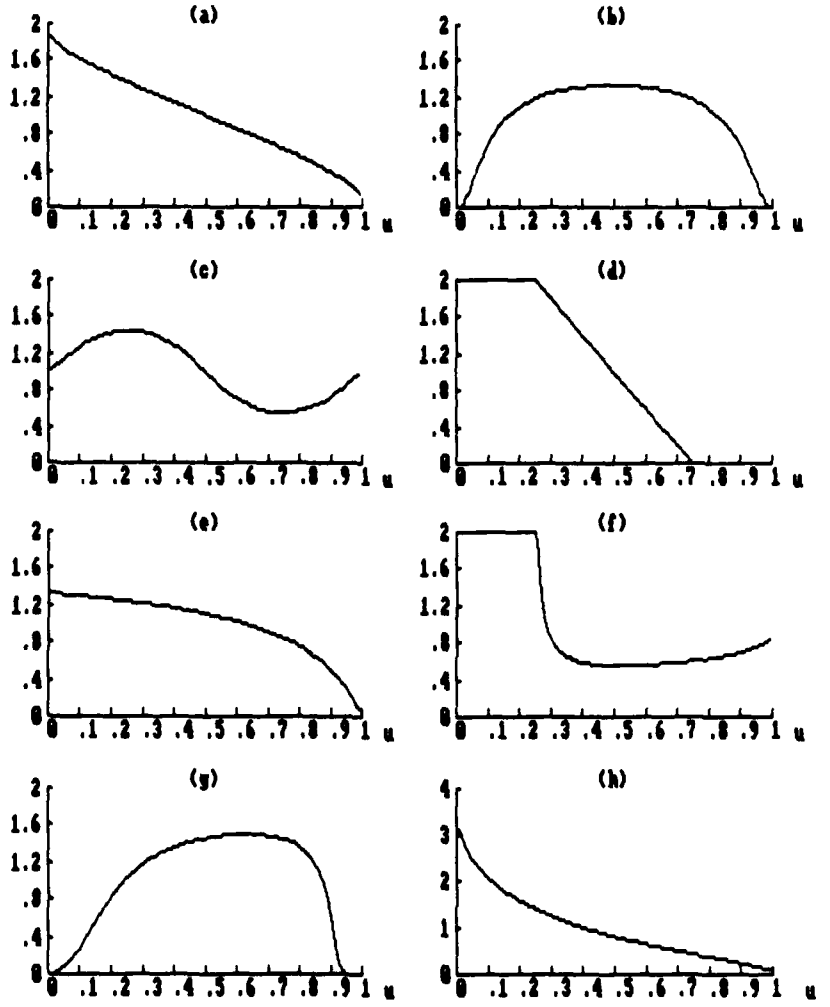

Fig. 1. *Examples of the comparison density function. Figure* (a) *is constructed using* $F = \mathrm{N}(0,1)$ *and* $G = \mathrm{N}(1,1)$*; Figure* (b)*,* $F = \mathrm{N}(0,1)$ *and* $G = \mathrm{N}(0,4)$*; Figure* (c)*,* $F = \mathrm{Cauchy}(0,1)$ *and* $G = \mathrm{Cauchy}(1,1)$*; Figure* (d)*,* $F = \mathrm{Triangular}(0,1)$ *and* $G = \mathrm{Triangular}(1,1)$*; Figure* (e)*,* $F = \mathrm{Exp}(1)$ *and* $G = \mathrm{Exp}(1/2)$*; Figure* (f)*,* $F = \mathrm{Lognormal}(0,1)$ *and* $G = \mathrm{Lognormal}(1,1)$*; Figure* (g)*,* $F = \mathrm{Weibull}(3)$ *and* $G = \mathrm{Exp}(1)$*; and Figure* (h)*,* $F = \mathrm{N}(0,1)$ *and* $G = \mathrm{N}(1,1)$*. Figures* (a) *through* (g) *are constructed with* $\lambda = 1/2$ *and Figure* (h) *is constructed using* $\lambda = 1/4$*.*

appears similar to Figure (a). One could define a pure location alternative as those alternatives for which d_λ is monotone and a pure scale alternative as those for which d_λ has one sign change of its derivative. However, Figures (c) and (f) fail such a criterion. It would seem preferable to classify as a location alternative those d_λ for which the dominant term in an orthonormal expansion of d_λ is the lowest frequency. Scale alternatives would have as a dominant term the next higher frequency. Figures (b) and (g) are quite similar; yet the first is a scale alternative and the second a general alternative. One could argue that in the case of Figure (g) that the dominant difference between the two is scale. In Section 3, diagnostics will be introduced which help indicate the types of relations of F to G present.

The comparison density is a specific example of the general technique of reducing the null hypothesis to a test of uniformity of a function defined on $[0, 1]$. The idea is well established in terms of the one sample problem in which the null hypothesis is completely specified. In the one sample location-scale problem, Parzen (1979) introduces the more general approach taken here. He terms this approach the density estimation approach to goodness of fit. In his comments on the article, Lindley remarks that this approach "provides something which looks as if it will be easier to handle than the raw functions."

The comparison density is seen as a starting point to fulfilling the criteria for testing outlined in Subsection 2.2.6. The comparison density, being a function, is graphical in nature and conveys information as to the relation of f to g. The goal is to estimate the comparison density, in a manner to be determined, and to test that estimate for uniformity. If the test rejects, a graph of the estimate is displayed and various diagnostics presented.

2.3.3. *The Comparison Distribution Empirical Process.* It has been seen that the comparison density, d_λ, is a useful and interesting object. It is desired both to estimate the comparison density and to derive inferential procedures concerning its uniformity. As pointed out in Subsection 2.2.6, however, these two goals are dual in nature, each depending on the other.

As a practical matter, one needs to determine a stochastic process such

that an estimator of d_λ can be written as a functional of this process. In this subsection, such a stochastic process is discussed. This stochastic process is termed the comparison distribution empirical process and is introduced by Parzen (1983) as a unifying concept. It will be seen to be a unification in that linear rank statistics, the Cramér-von Mises, Anderson-Darling and Kolmogorov-Smirnov statistics can all be conveniently represented as functionals of this process.

As a means of motivating this approach, consider estimating the mean, μ, of the distribution, F, from the first sample, $X_1, \ldots, X_m$. Suppose that F has finite variance σ^2. The sample mean, $\bar{X}$, is given by $\bar{X} = \int x dF_m(x)$, where F_m is the empirical distribution function. One is led naturally to the statistic

$$\begin{aligned}\sqrt{m}(\bar{X} - \mu) &= \sqrt{m} \int x d[F_m(x) - F(x)] \\ &= \int_0^1 Q^F(t) d[F_m Q^F(t) - t] \\ &= \int_0^1 Q^F(t) dU_m(t) \\ &= \int_0^1 q^F(t) U_m(t) dt,\end{aligned}$$

where $q^F(t) = \frac{d}{dt} Q^F(t)$. The quantity $U_m(t)$ is termed the uniform empirical process [see Shorack and Wellner (1986), page 86] and it is well known [Shorack and Wellner, page 110] that $U_m(t) \Rightarrow U(t)$ as $m \to \infty$, where $\Rightarrow$ denotes weak convergence. Here, $U(t)$ is a Brownian bridge; it can be characterized as $U(t) = W(t) - tW(1)$, where $W(t)$ is a Wiener process. The Wiener process is defined as a process that satisfies $W(0) = 0$ a.s., $W(t) \sim N(0, t^2)$, and $W(t)$ has stationary, independent increments. Hence, the Brownian bridge is a Gaussian process with mean function $\mathrm{E}[U(t)] = 0$, covariance kernel $K(s,t) = \mathrm{E}[U(t)U(s)] = \min(s,t) - st$, and $U(0) = U(1) = 0$ a.s.. See Billingsley (1986), page 522, or Shorack and Wellner (1986) for more details.

Given the weak convergence of U_m to U, one would hope that it would follow that

$$\int_0^1 q^F(t) U_m(t) dt \xrightarrow{\mathrm{d}} \int_0^1 q^F(t) U(t) dt \text{ as } m \to \infty.$$

This last integral is normally distributed with mean

$$\mathrm{E}\Big[\int_0^1 q^F(t)U(t)dt\Big] = \int_0^1 q^F(t)\mathrm{E}[U(t)]dt = 0,$$

and variance

$$\begin{aligned}\mathrm{Var}\Big[\int_0^1 q^F(t)U(t)dt\Big] &= \int_0^1 QF(t)^2dt - \Big[\int_0^1 QF(t)dt\Big]^2 \\ &= \int x^2 f(x)dx - \Big[\int xf(x)dx\Big]^2 \\ &= \sigma^2,\end{aligned}$$

[See Parzen (1962b), page 77, and Parzen (1979)]. It has just been proved, albeit somewhat heuristically, that $\bar{X}$ is AN(μ,σ^2/m). The key is to define an empirical process such that the parameter one is interested in can be written as a functional of that process. For this example, one need only apply the central limit theorem to achieve the result. The case of estimating d_λ is far less straightforward in that one is not estimating a single parameter from a random sample. The stochastic process approach is the only one available.

The parameter λ represents the weight or proportion given the distribution F. For the samples $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$, the natural estimate of λ is $\lambda_{(N)} = m/N$, where $N = m + n$. It is assumed throughout that $\lambda_{(N)} \to \lambda_0$ as $m \wedge n \to \infty$, where $0 < \lambda_0 < 1$. Define $D_{(N)}(w) = D_{\lambda_{(N)}}(w)$, $d_{(N)}(w) = d_{\lambda_{(N)}}(w)$, $D_0(w) = D_{\lambda_0}(w)$, and $d_0(w) = d_{\lambda_0}(w)$. These first two functions depend on N, but only through the parameter $\lambda_{(N)}$, and so are not random. In fact, the goal is the estimation of $d_{(N)}$ which is tantamount to estimating d_0 as $m \wedge n \to \infty$.

The obvious process to define for the comparison distribution function is

$$L_N(w) = \sqrt{N}[K_N(w) - D_{(N)}(w)], \tag{2.3.3}$$

where $K_N(w) = F_m Q_N^H(w)$, and Q_N^H is the sample quantile function of the pooled sample. The process (2.3.3) simply substitutes the empirical functions for the unknowns, which is the manner in which U_m, above, was constructed. Let $X_{(i)}$ be the i^{th} order statistic of $X_1,\ldots,X_m$, for $i = 1,\ldots,m$. Let R_i be the rank of

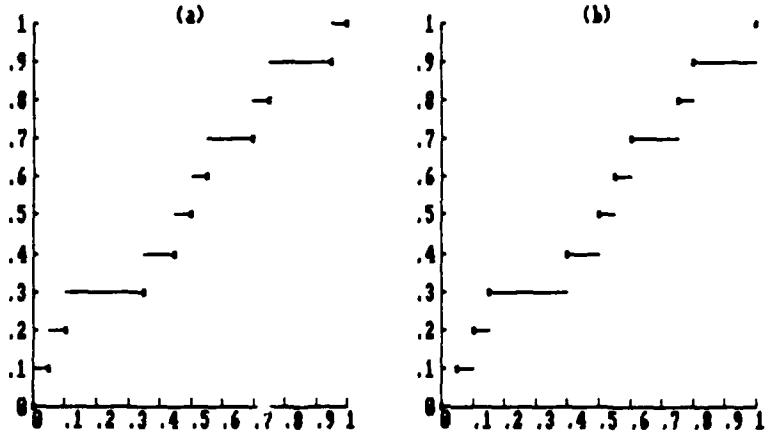

Fig. 2. *The typical appearance of K_N and D_N. Figure (a) is K_N and (b) is D_N. The two graphs are constructed using the same input ranks and $m = 10$ and $n = 10$. The block on the end of the line segments marks the value of the function at the jump points.*

X_i in the pooled sample and let $R_{(i)}$ be the rank of $X_{(i)}$ in the pooled sample. This latter notation is appropriate since $R_{(i)}$ is also the i^{th} order statistic of $R_1, \ldots, R_m$. Now K_N is given by

$$K_N(w) = \begin{cases} 0, & \text{for } 0 \le w \le (R_{(1)} - 1)/N; \\ j/m, & \text{for } (R_{(j)} - 1)/N < w \le (R_{(j+1)} - 1)/N \text{ and } j = 1, \ldots, m; \\ 1, & \text{for } (R_{(m)} - 1)/N < w \le 1. \end{cases}$$

Figure 2(a) gives the typical appearance of K_N.

Pyke and Shorack (1968) study the process $L_N(w)$ extensively. Their main result is that under conditions (2.3.1), $L_N(w) \Rightarrow L(w)$ as $m \wedge n \to \infty$, where

$$L(w) = (1 - \lambda_0)\Big(d_0^G(w)U[D_0(w)]/\sqrt{\lambda_0} - d_0(w)V[D_0^G(w)]/\sqrt{1 - \lambda_0}\Big), \tag{2.3.4}$$

and U, V are independent Brownian bridges and D_0^G and d_0^G satisfy $\lambda_0 D_0^G(w) + (1 - \lambda_0)D_0(w) = w$ and $\lambda_0 d_0^G + (1 - \lambda_0)d_0 = 1$. The process $L(w)$ is Gaussian with mean function 0 and covariance kernel $K(u, v) = \mathrm{E}[L(u)L(v)]$, equal to

$$K(u,v) = (1 - \lambda_0)^2\Big(d_0^G(u)d_0^G(v)D_0(u)[1 - D_0(v)]/\lambda_0 + d_0(u)d_0(v)D_0^G(u)[1 - D_0^G(v)]/(1 - \lambda_0)\Big), \tag{2.3.5}$$

for $u < v$. If $F = G$, $K(u,v)$ and $L(w)$ simplify tremendously and one finds that $L(w) = (1-\lambda_0)[U(w)/\sqrt{\lambda_0} - V(w)/\sqrt{1-\lambda_0}]$ and $K(u,v) = (1-\lambda_0)u(1-v)/\lambda_0$ for $u < v$.

Parzen (1983) chooses to define a slightly different process, $\mathrm{CD}_N(w)$, as

$$\mathrm{CD}_N(w) = \sqrt{N}[D_N(w) - D_{(N)}(w)],$$

where $D_N(w) = [H_N Q_m^F]^{-1}(w)$; H_N is the empirical distribution function of the pooled sample; Q_m^F is the empirical quantile function of the first sample; and the exponent -1 refers to a special type of inverse which is given below. The function $H_N Q_m^F(u)$ has values,

$$H_N Q_m^F(u) = \begin{cases} 0, & \text{for } u = 0; \\ R_{(j)}/N, & \text{for } (j-1)/m < u \le j/m \text{ and } j = 1,\ldots,m; \end{cases}$$

since $Q_m^F = X_{(j)}$ for $(j-1)/m < u \le j/m$. Notice that $H_N Q_m^F$ is defined on 0 to 1, is non-decreasing and left continuous. These are the characteristic properties of a quantile function. Its inverse is defined as $D_N(w) = \sup\{u : H_N Q_m^F(u) \le w\}$ and results in

$$D_N(w) = \begin{cases} 0, & \text{for } 0 \le w < R_{(1)}/N; \\ j/m, & \text{for } R_{(j)}/N \le w < R_{(j+1)}/N \text{ and } j = 1,\ldots,m-1; \\ 1, & \text{for } R_{(m)}/N \le w \le 1. \end{cases}$$

As a matter of notation, D_N is a stochastic process; it is estimating $D_{(N)}$ which is not random. If there are parentheses around the N in a subscript, that quantity is not random. If there are no parentheses, the quantity is random. Note that $D_N(w)$ is non-decreasing, right continuous, and $D_N(0) = 0$ and $D_N(1) = 1$. These are the characteristic properties of a distribution function on $[0,1]$. Figure 2(b) presents the function D_N for the same data as is used to construct L_N in Figure 2(a). Aly, Csörgő, and Horváth (1987) use embedding techniques to prove that $\mathrm{CD}_N(w)$ converges weakly to the process $L(w)$. Parzen (1983) conjectures that this result can be obtained from Pyke and Shorack's (1968) results since $D_N(w)$ and $K_N(w)$ differ by $1/m$ on m intervals of length $1/N$. A proof is as follows. The function $D_N(w)$ can be written as $D_N(w) = L_N(w) + \Delta_N(w)$, where

$$\Delta_N(w) = \begin{cases} 1/m, & \text{for } (R_{(j)} - 1)/N < w < R_{(j)}/N \text{ and } j = 1, \ldots, m; \\ 0, & \text{otherwise.} \end{cases}$$

Pyke and Shorack (1968) give the following representation for $L_N(w)$:

$$(2.3.6) \quad L_N(w) = (1 - \lambda_{(N)})\Big(B_N(w)U_m[FQ_N^H(w)]/\sqrt{\lambda_{(N)}} - A_N(w)V_n[GQ_N^H(w)]/\sqrt{1 - \lambda_{(N)}}\Big) + \delta_N(w) \text{ a.s.},$$

where $A_N(w) = [D_N(u_w) - D_N(w)]/(u_w - u)$ and $u_w = HQ_N^H(w)$; $B_N(w)$ and δ_N are defined by $\lambda_{(N)}A_N(w) + (1 - \lambda_{(N)})B_N(w) = 1$ and $\delta_N(w) = A_N(w)\sqrt{N}(H_NQ_N^H(w) - w)$; U_m and V_n are the uniform empirical processes for the first and second sample, respectively. Pyke and Shorack use the Skorohod device [see Shorack and Wellner (1986), page 54] to create versions, $\tilde{U}_m$, $\tilde{V}_n$, $\tilde{U}$, and $\tilde{V}$ of U_m, V_n, U, and V such that the new versions are distributed identically as the original versions, are defined on a common probability space, and satisfy

$$\|\tilde{U}_m - U\| \to 0 \text{ a.s.},$$
$$\|\tilde{V}_n - V\| \to 0 \text{ a.s.},$$

as $m \wedge n \to \infty$, where $\|\cdot\|$ is the sup-norm. If one can show convergence in probability for the new processes, this would imply that their probability measures also converge. Since these probability measures are identical to those of the original processes, this must mean that they converge also and hence that one has weak convergence. These new versions are substituted into equations (2.3.4) and (2.3.6) to obtain $\tilde{L}(w)$ and $\tilde{L}_N(w)$, respectively. They then show that

$$\|\tilde{L}_N(w) - \tilde{L}(w)\| \to_p 0 \text{ as } m \wedge n \to \infty.$$

Write $\tilde{D}_N(w) = \tilde{L}_N(w) + \Delta_N(w)$ so that

$$\begin{aligned} \|\tilde{D}_N(w) - \tilde{L}(w)\| &= \|\tilde{L}_N(w) + \Delta_N(w) - \tilde{L}(w)\| \\ &\leq \|\tilde{L}_N(w) - \tilde{L}(w)\| + \|\Delta(w)\| \\ &\to_p 0 \text{ as } m \wedge n \to \infty. \end{aligned}$$

Hence, the same proof works with little additional work. It is important to establish this fact as weak convergence under local alternatives will be shown for $L_N(w)$ in Section 4 and the result carried over to $D_N(w)$.

The stochastic process suggested by Parzen will be the basic process used in this work. It is chosen over that of Pyke and Shorack for several reasons. First, $D_N(w)$ has the form of the sample distribution function constructed from the data $R_1/N, \ldots, R_m/N$. This sample distribution function is estimating the true distribution function whose density it is desired to estimate. The analogy with the ordinary density estimation is very strong as shall be seen in Subsection 2.4. In this case, one views $R_1/N, \ldots, R_m/N$ as data arising from the density $d_{(N)}$ and uses conventional density estimators for $d_{(N)}$. As with the example of the sample mean, these estimators can be written as functionals of a stochastic process. The limiting distribution of these functionals differs from the usual case because the limiting distribution of the underlying stochastic process is different. Second, $D_N(w)$ is preferred because rank statistics are more easily represented as functionals of $D_N(w)$. Recall the rank statistic $S = \sum_{j=1}^{m} J(R_i/N)$ as defined by (2.2.1). This statistic can be neatly rewritten as

$$S_N = \int_0^1 J(w) dD_N(w).$$

In fact, one can rewrite the centered form of the statistic as

$$\begin{aligned} S_N^* &= \int_0^1 J(w) d\mathrm{CD}_N(w) \\ &= \int_0^1 J'(w) \mathrm{CD}_N(w) dw. \end{aligned}$$

The asymptotic normality of S_N^* is almost trivial if J is differentiable on $(0,1)$ and the derivative is bounded there. In this case, the functional $K(f) = \int_0^1 J'(w) f(w) dw$ is uniformly continuous for $f \in D[0,1]$, the set of all functions on $[0,1]$ which have limits from the left and are continuous from the right. By Theorem 3.12 of Ruymgaart (1988), S_N^* converges in distribution to $\int_0^1 J'(w) L(w) dw$, which is a Gaussian random variable. Aly *et al.* (1987) study more general score functions which are allowed to depend on N. The representation for rank statis-

tics used by Pyke and Shorack (1968) is

$$S_N^* = \int_0^1 L_N(w) d\nu_N,$$

where ν_N is a signed measure which puts measure c_{Ni} on the point i/N for $i = 1, \ldots, N$. This representation is somewhat cumbersome and less natural than that derived for the CD_N process.

2.3.4. *Existing Work on the Estimation of the Comparison Density.* In this subsection, existing work on estimating and testing the uniformity of the comparison density is reviewed. Some very interesting work has been done in the area, but it is seen that these techniques fall short of the goals which have been outlined.

Parzen (1983) derives a testing and estimation methodology which fits into the framework outlined in Subsection 2.2.6. His approach is essentially the same that will be taken here: to apply a general method for the estimation of densities to the special stochastic process CD_N. Parzen's estimator is known as the autoregressive estimator and its use in the general density estimation setting is detailed in Parzen (1979) and Carmichael (1984). It is also discussed in Subsection 2.4.5.

The estimate of the comparison density, $\hat{d}_k$, is defined by

$$\hat{d}_k(w) = \sigma_m^2 \left| 1 + \sum_{j=1}^{k} \alpha_j e^{2\pi i j w} \right|^{-2},$$

where the α_j's are complex-valued and $|\cdot|^2$ denotes the complex squared modulus. The parameter, k, is a smoothing parameter and is referred to as the order of the autoregressive process. Larger values of k lead to rougher estimates. The α_j's and σ_m^2 are estimated from the data, $R_1/N, \ldots, R_m/N$. The form of $\hat{d}_k(w)$ is that of the spectral density [see Newton (1988)] associated with a complex-valued $\mathrm{AR}(k)$ process with coefficients, $\alpha_1, \ldots, \alpha_k$, and normalized residual variance, σ_k^2; hence the name.

Parzen (1983) defines the pseudo-correlations to be

$$\hat{\rho}_k = \int_0^1 e^{2\pi i k x} dD_N(x), \text{ for } k = 0, 1, 2, \ldots, .$$

The estimates of $\alpha_j, j = 1, \ldots, k$ and σ_k^2 are $\hat{\alpha}_j, j = 1, \ldots, k$ and $\hat{\sigma}_k^2$, respectively. They may be obtained utilizing a complex-valued version of Levinson's algorithm (see Parzen (1983) for a description of the algorithm) based on the pseudo-correlations. Parzen suggests choosing k by a version of Akaike's (1974) AIC criterion, which chooses k to minimize

$$\mathrm{AIC}(k) = \ln \hat{\sigma}_k^2 + \frac{2k}{N} \frac{1 - \lambda_{(N)}}{\lambda_{(N)}}$$

if that value of AIC is negative and selects k to be zero otherwise. The selection of $k = 0$ is significant because $\hat{d}(w) = 1$ if $k = 0$. The AIC selection criterion can then also be viewed as a test of the null hypothesis, $\mathrm{H}_o: d_{(N)}(w) = 1$. If AIC chooses $k = 0$, the null hypothesis is accepted. If AIC chooses $k > 0$, the null hypothesis is rejected and a model is chosen. This simultaneous testing and model selection is exactly what is being sought here. However, neither the autoregressive estimator nor AIC will be used in this dissertation. The procedure, however, stands as a benchmark to which to compare any new procedures.

There are several difficulties with this procedure both as a test of H_o and as an estimator. As a test, the properties of AIC in this framework are unknown. In particular, the size of this test (the probability of rejection if H_o is true) is unknown. Nor are there any provisions for adjusting the size to a pre-specified level. These two are not damning criticisms. Although one would probably not expect to solve them analytically, they certainly would yield to simulation techniques. Of much more concern is the behavior of $\hat{d}_k$ as an estimator. It can be shown that $\hat{d}_k$ satisfies the relation

$$\hat{d}_k(0) = \hat{d}_k(1).$$

Such a condition is referred to as a periodicity condition. There is no reason to suppose that $d_\lambda(u)$ satisfies such a condition. If it does not, the estimator is biased and inconsistent at the ends. It will be seen in Subsection 2.4 that such biases can reduce the efficiency of an estimator drastically.

Eubank, LaRiccia, and Rosenstein (1987) investigate what they term the components of Pearson's phi-squared distance measure. Pearson's phi-squared,

as defined by Eubank *et al.* (1987), is

$$\phi^2 = \int_0^1 (d_{(N)}(w) - 1)^2 dw,$$

which is the squared $\mathcal{L}_2[0,1]$ norm between $d_{(N)}$ and 1. Certainly $H_o: d_{(N)}(w) = 1$, $0 \leq w \leq 1$ is equivalent to $H_o: \phi^2 = 0$. Unfortunately, there are no natural estimators for ϕ^2, although an estimate of the form

$$\int_0^1 (\hat{d}(w) - 1)^2 dw$$

is investigated in Section 3. Eubank *et al.* suggest instead decomposing ϕ^2 into its components. Start by selecting a complete orthonormal sequence for $\mathcal{L}_2[0,1]$ (see Subsection 2.4.4), $\{p_j(w)\}$, and define

$$a_j = \int_0^1 [d_{(N)}(w) - 1] p_j(w) dw \text{ for } j = 1, 2, \ldots, .$$

The a_j's are the components of Pearson's phi-squared and they satisfy

$$\phi^2 = \sum_{j=1}^{\infty} a_j^2.$$

The components are estimated by

$$\hat{a}_j = \int_0^1 p_j(w) d\mathrm{CD}_N(w). \tag{2.3.7}$$

These components bear a marked similarity to the components of the Anderson-Darling and Cramér-von Mises statistics discussed in Subsection 2.2.5. In point of fact, choosing $p_j(w) = \sin \pi j w$ yields the components of the Cramér-von Mises statistic and $p_j(w)$ as the Legendre polynomials yields the components of the Anderson-Darling statistic.

A test of $H_o: \phi^2 = 0$ is then equivalent to $H_o: a_j = 0$ for $j \geq 1$. Of course, this latter hypothesis is not testable. One cannot simultaneously test an infinite number of parameters. One could weight the components by forming a statistic like $\sum \lambda_j \hat{a}_j^2$ where $\sum \lambda_j^2 \mathrm{Var}(\hat{a}_j) < \infty$ to arrive at an asymptotically consistent test. Recall this is the form of the Anderson-Darling and the Cramér-von Mises statistics. Eubank *et al.* suggest instead that one test subhypotheses, such as

$$H_o: a_j = 0 \text{ for } j = 1, \ldots, M \tag{2.3.8}$$

This latter suggestion is most intriguing as one then gives equal weight in the testing procedure to each of the first M a_j's. This notion will be discussed at greater length in Section 3. In terms of implementing this suggestion, note that under H_o

$$\begin{aligned}\mathrm{Cov}(\hat{a}_j, \hat{a}_k) &= \frac{1-\lambda_{(N)}}{\lambda_{(N)}} \int_0^1 \int_0^1 p_j'(u) K(u,v) p_k'(v) du dv \\ &= \frac{1-\lambda_{(N)}}{\lambda_{(N)}} \Big[\int_0^1 p_j(w) p_k(w) dw - \int_0^1 p_j(w) dw \int_0^1 p_k(w) dw \Big] \\ &= -\frac{1-\lambda_{(N)}}{\lambda_{(N)}} \int_0^1 p_j(w) dw \int_0^1 p_k(w) dw.\end{aligned}$$

In particular, if the orthonormal sequence, $\{p_j(w)\}$, is also orthogonal to $p_0(w) = 1$ then the components are asymptotically Gaussian and independent with variance

$$\frac{1-\lambda_{(N)}}{\lambda_{(N)}} \int_0^1 p_j^2(w) dw.$$

From the form of (2.3.7), it is seen that the components are also rank statistics. This interpretation of the components brings to light several interesting prospects. First, Eubank *et al.* (1987) note that the usual form of the sequence of $p_j(w)$'s is that they become more oscillatory as j increases. Typically, p_1 will have one zero crossing in $(0,1)$, p_2 will have two, p_3 three, and so on. A score function with one zero crossing is testing location; one with two crossings tests scale. More crossings can be viewed as testing higher frequency departures from uniformity. Testing the hypothesis (2.3.8) for the first 2 components then results in a test against both location and scale. The independence of location and scale rank statistics is known in the literature [see Randles and Hogg (1971) and Boos (1986)], but it is presented in an *ad hoc* fashion with no unifying philosophy. Eubank *et al.* seem to be the first to give any framework and extension to this observation. One might choose the parameter, M, based on how great a departure is deemed worthy of testing.

Second, one can choose the orthonormal sequence so that it protects best against certain distributions or types of tail behavior. For example, if one wished

to design a test to be most powerful in the case when F and G are long tailed, one might choose to use the cosine functions for the sequence. The cosine is the optimal score function of the Cauchy. Similarly, the Legendre polynomials would be suitable for medium tailed distributions, such as the logistic distribution.

Although they do not investigate the properties at all, Eubank *et al.* (1987) note that the components do lead to an estimate of $d_{(N)}(w)$ as, say,

$$\hat{d}(w) = \sum_{j=1}^{M} \hat{a}_j p_j(w).$$

Such estimates are referred to as orthogonal series estimates and will be discussed in Subsection 2.4.4. Eubank *et al.* present some very intriguing ideas, several of which will be taken up in later sections. However, they do not outline a comprehensive testing and estimation procedure that is being sought here.

2.4. Review of Density Estimation Techniques on [0,1]

2.4.1. *Introduction.* The purpose of this subsection is to review various methods that can be used to estimate $d_\lambda(u)$. These techniques fall under the general heading of density estimation. Since $d_\lambda(u)$ is known to have support [0,1], the implications of this fact on the properties of the estimators must be closely examined. Modifications of certain estimators for this case, particularly kernel estimators, have been proposed in the literature. These, too, will be reviewed.

As with two sample tests, density estimators can also be classified as parametric or nonparametric. A parametric estimator of a density assumes that $f \in \mathcal{F}_\theta$ where the family of distributions, $\mathcal{F}_\theta$, is defined as $\mathcal{F}_\theta = \{f(x) = f(x;\theta), \theta \in \Theta, \Theta \subseteq \mathbb{R}^k\}$. Given a random sample, $X_1, \ldots, X_n$, from $f(\cdot;\theta)$, θ is estimated by a method such as maximum likelihood, minimum chi-square, or the method of moments [see Kendall and Stuart (1979)]. The resultant estimator of f is $\hat{f} = f(\cdot;\hat{\theta})$. In particular, if $\hat{\theta}$ is the maximum likelihood estimator of θ then $f(\cdot;\hat{\theta})$ is the maximum likelihood estimator of f by the invariance principle [see Mood, Graybill, and Boes (1974), page 284].

Philosophically, should there be justification to assume that the underlying density belongs to a parametric family it is, as Good and Gaskins (1971) remark, "a pity to waste it." It will be seen that parametric estimators generally enjoy faster rates of convergence of mean squared error (MSE) than nonparametric ones. In the case at hand, namely estimation of $d_\lambda(u)$ from $R_1/N, \ldots, R_m/N$, parametric techniques simply do not apply. There is no reason to suspect that the class $\mathcal{D}$ of all $d_\lambda(u)$ constructed as in Subsection 2.3 can be indexed by a finite dimensional parameter, θ. For this reason, it is necessary to look into the realm of nonparametric techniques of density estimation for a methodology.

Several nonparametric density estimators will be examined in detail. These are the histogram, the kernel method, the orthogonal series method and AR/ARMA methods. In this discussion, it is assumed that one desires to estimate a density, f, and has at hand a random sample, $X_1, \ldots, X_n$, from f. The discussion will be organized as follows. First the estimator is defined and its properties given. Special attention will be given to representations of mean squared error and mean integrated squared error (MISE). Other properties such as weak and strong consistency will be referenced but not detailed. Second, the implications of f having support $[0,1]$ are examined. If modifications to the original estimator have been proposed, these will be discussed.

It is assumed that the reader is familiar with the basic concepts of nonparametric density estimation and related standard terms such as MISE. Background material can be found in the following books and review articles: Silverman (1986), Titterington (1985), Bean and Tsokos (1980), Tapia and Thompson (1978), Wertz (1978), Scott, Tapia, and Thompson (1977), Wegman (1972a) and Wegman (1972b).

At this point it seems wise to reiterate that the properties of the estimators described in the following subsections are derived for data, $X_1, \ldots, X_n$, which is iid f. Since this is not the case for $R_1/N, \ldots, R_m/N$ it should not be expected that the properties should carry over in a one to one fashion. Since the iid case is in many ways ideal, such results may indicate the best that can be done.

2.4.2. *The Histogram.* The histogram is the oldest of the nonparametric

density estimators and is best suited to f having compact support. In fact, difficulties arise should f have infinite support. Despite this, it is seen that the histogram is not best suited to the needs at hand.

The histogram is constructed for data, $x_1, \ldots, x_m$, in the following manner. Select bin edges $t = (t_0, \ldots, t_m)$ such that $0 = t_0 < t_1 < \cdots < t_m = 1$. The histogram estimate is given by

$$f^h(x;t) = \frac{n_i}{n(t_i - t_{i-1})}, \text{ for } t_{i-1} \leq x < t_i,$$

where n_i is the number of data points falling in the interval $[t_{i-1}, t_i)$.

Two of the simplest properties of $f^h(x;t)$ are easily verified; namely $f^h(x;t) \geq 0, \forall x$ and $\int f^h(x;t)dx = 1$. These imply the estimate is itself a probability density function. Tapia and Thompson (1978) prove the following theorem concerning the consistency in mean square of the histogram.

Theorem 2.4.1 (Tapia and Thompson, 1978). *Suppose that f has continuous derivatives up to order three except at the endpoints of $[0,1]$, and f is bounded on $[0,1]$. Let the mesh be equal spacing throughout $[0,1]$, so that $t_i - t_{i-1} = 2h$. If $h \to 0$ and $nh \to \infty$ as $n \to \infty$, (note the partition is now a function of n) then for $x \in (0,1)$, $\mathrm{MSE}_x(f^h, f) \to 0$, where*

$$\mathrm{MSE}_x(f^h, f) = \mathrm{E}\Big([f^h(x;t) - f(x)]^2\Big).$$

From the details of the proof of Theorem 2.4.1, an upper bound on MISE can be derived. This bound is

$$\mathrm{MISE}(f^h, f) \leq \frac{1}{nh} + 2h^2 \int_0^1 f'(x)^2 dx + O(n^{-1}) + O(h^3).$$

Minimizing this bound with respect to the bin width, h, one finds that the best rate of convergence of $\mathrm{MISE}(f^h, f)$ is $O(n^{-2/3})$.

There are several criticisms of the histogram. The first is the rate of convergence of MISE. The kernel estimators examined in the next subsection do better than $O(n^{-2/3})$. Second, it seems unfortunate to estimate a function which has been assumed to possess three continuous derivatives by a step function. If f is smooth, it is desirable that the estimate should be, too.

2.4.3. *The Kernel Method.* This subsection reviews kernel density estimators. With some modification for boundary effects, it is seen that kernels are a viable method of density estimation on $[0,1]$. Rosenblatt (1956) first suggests the method of kernel density estimation. He examines in detail only the rectangular kernel, however. Parzen (1962a) investigates general kernels and derives myriad results. In fact, such has been the influence of his research in this area that they are often referred to as Parzen kernel estimators.

The kernel density estimator, f^k, is defined as

$$f^k(x;h) = \frac{1}{h}\int K\left(\frac{x-y}{h}\right) dF_n(y) \tag{2.4.1}$$
$$= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right),$$

where F_n is the empirical distribution function and h is the bandwidth or window width. Parzen establishes the following theorem concerning the mean square consistency of f^k.

Theorem 2.4.2 (Parzen, 1962a). *Let x be a continuity point of f and suppose $h \to 0$ and $nh \to \infty$ as $n \to \infty$. Assume that K is bounded, absolutely integrable, and $|yK(y)| \to 0$ as $y \to \infty$; then $\mathrm{MSE}_x(f^k, f) \to 0$ as $n \to \infty$.*

Under various additional assumptions on h, K and f, he also establishes asymptotic normality and uniform weak consistency.

Strong consistency has been considered by several authors: Silverman (1978), Bertrand-Retali (1978) and Nadarya (1965). Wahba (1975) derives minimax results for the MSE of kernel estimators. That is, for suitable restrictions on K and h and $f \in W$, where W is an appropriate space of densities, Wahba derives an upper bound on $\mathrm{MSE}_x(f^k, f)$ for all $f \in W$.

Parzen (1962a) gives an asymptotic representation of MSE and MISE of f^k. He assumes the existence of an integer $r \geq 0$ such that

$$\kappa_r = \lim_{u \to 0} \frac{1 - k(u)}{|u|^r},$$

is finite and nonzero where $k(u)$ is the Fourier transform of K. The number κ_r is called the characteristic coefficient and r the characteristic exponent of the kernel

K. Parzen (1962a) assumes also that $f^{(r)}(x) = \int e^{-iux}|u|^r\varphi(u)du$ converges absolutely, where $f^{(r)}$ is the r^{th} derivative of f and φ is the characteristic function of f. In this case, MISE and MSE admit the following expansions:

$$\text{(2.4.2)} \qquad \text{MSE}_x(f^k, f) \sim \frac{f(x)}{nh}\int K(y)^2dy + h^{2r}\left|\kappa_r f^{(r)}(x)\right|^2,$$

$$\text{(2.4.3)} \qquad \text{MISE}(f^k, f) \sim \frac{1}{nh}\int K(y)^2dy + h^{2r}k_r^2\int f^{(r)}(x)^2dx.$$

In each of these expansions, the first term is the contribution due to variance and the second due to squared bias. Minimizing (2.4.3) with respect to h, one sees that the best rate of convergence of MISE is $O(n^{-2r/(2r+1)})$. K is normally chosen to be a probability density function which is symmetric about 0 and has a finite variance (which implies $r = 2$), in which case the best rate of convergence is $O(n^{-4/5})$. This rate is better than that of the histogram.

The discussion will now center on events at $x = 0$ (however, any conclusions hold for the other endpoint, $x = 1$). Suppose that it is known that f has support $[0, 1]$ and is continuous on $(0, 1)$. If $f(0) = f(1) = 0$, so that f is continuous on $\mathbb{R}$, then all the standard results noted above still apply. Now suppose that $f(0) > 0$; f has a simple discontinuity at $x = 0$. Theorem 2.4.2 now fails at $x = 0$.

Let K be a symmetric density function and consider $f^k(0; h)$. Equation (2.4.1) simplifies to

$$f^k(0; h) = \frac{1}{h}\int_0^1 K(-y/h)dF_n(y),$$

which means that one is using $K(y)$ only for $y \leq 0$. In this case, one can define the effective kernel to be $K^e(y) = K(y)I_{(-\infty,0]}(y)$, where I is the indicator function. For K^e the characteristic exponent is $r = 0$ since $\int K^e(y)dy = \frac{1}{2}$. Referring to equation (2.4.2), one sees that $\text{MSE}_0(f^k, f) \to \frac{1}{4}f(0)^2$ as $h \to 0$, $nh \to \infty$ and $n \to \infty$, or that $\text{Bias}_0(f^k, f) \to -\frac{1}{2}f(0)$. The problem is that the f^k is converging in mean square to $\frac{1}{2}f(0)$ [see Schuster (1985)]. There is no difficulty with the variance term, only the bias term.

To investigate this phenomenon more closely, start by assuming that K has compact support. For $x \geq h$, there is no problem and the usual definition (2.4.1)

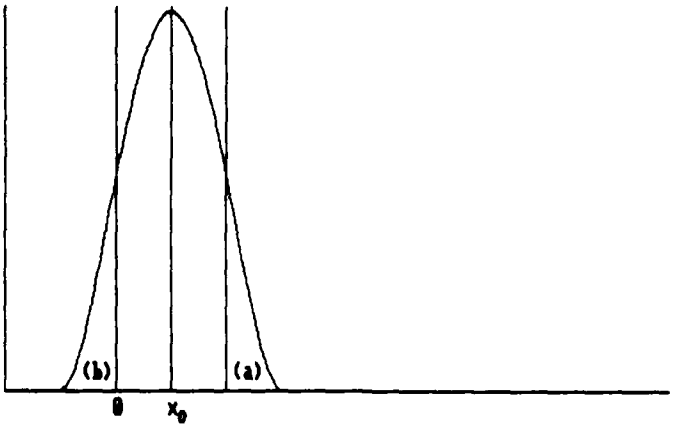

Fig. 3. *Position of the kernel to estimate the density at x_0. Region (a) on the right corresponds to region (b) on the left. Region (a) is unused in the estimation.*

applies. For $x < h$, part of the kernel is clipped; it will never be used since f is zero on $(-\infty, 0)$. So $f^k(x;h)$ can be rewritten as

$$f^k(x;h) = \frac{1}{h}\int_0^1 K_s^e\left(\frac{x-y}{h}\right) dF_n(y),$$

where $s = x/h$ and

$$K_s^e(y) = \begin{cases} K(y)I_{[-1,s]}(y), & \text{for } 0 \le s \le 1; \\ K(y)I_{[-1,1]}(y), & \text{for } s \ge 1, \end{cases}$$

is the effective kernel. Figure 3 depicts this phenomenon graphically. For estimating f at x_0, one uses a kernel "sitting" on the point x_0. A portion of the right tail of K, which is the mirror image of that portion of the left tail falling below 0, is never used. So, instead of using a kernel with $r = 2$, one is actually using a kernel with $r = 0$. The bias is greatly increased. However, for points away from $x = 0$, the MSE behaves as usual asymptotically. This is so because, for fixed x_0, $s = x_0/h = 1$ when $h = x_0$. As h decreases to x_0, the kernel sitting at x_0 no longer reaches to $x < 0$ and so the effective kernel, K^e, is just K. Since h is tending to 0, eventually $h = x_0$ for every $x_0 > 0$ and so the usual asymptotics apply to all $x > 0$.

From this discussion, it should be plain why only kernels of compact support are considered. Kernels of infinite support will always cross over to $x < 0$ and so

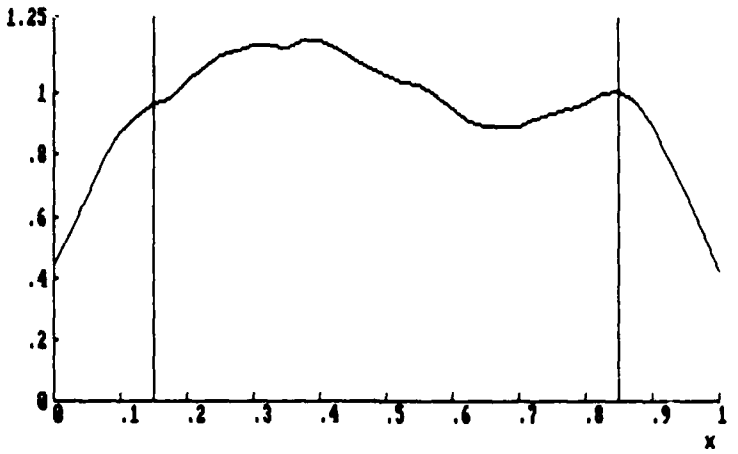

Fig. 4. *The kernel density estimate for a random sample of size* 250 *from the uniform* $[0,1]$ *distribution. The estimate is constructed using the biweight kernel and* $h = 0.15$.

will always be clipped. Hence, no matter how small h is taken, boundary effects will be experienced for each $x \in [0,1]$. Gasser and Müller (1979) make note of this fact.

In finite samples, one should expect to see boundary effects for x between 0 and h. Figure 4 demonstrates this point. The figure presents the kernel density estimate based on 250 iid observations taken from the uniform $[0,1]$ distribution. The biweight kernel, $K(t) = \frac{15}{16}(1-t^2)^2 I_{[-1,1]}(t)$, and bandwidth, $h = 0.15$, are used. Notice how the estimate starts to bend downward for $x < 0.15$ and $x > 0.85$. Clearly, the situation is not satisfactory.

Several proposals have been made to correct the situation. The cut and normalize method normalizes the effective kernel, K_s^e, so that it integrates to one. Define the cut and normalize kernel $K_s^{cn}(t)$ as

$$K_s^{cn}(t) = K_s^e(t) / \int_{-1}^{1} K_s^e(y)dy.$$

This normalization moves the characteristic exponent from $r = 0$ to $r = 1$ for $0 \le s < 1$. In the interior, one is using a kernel with characteristic exponent $r = 2$ and at the boundaries a kernel with characteristic exponent $r = 1$. To examine the MISE of this estimator, define $\mu(s) = \int_{-1}^{M(s)} K(t)^2 dt$ and $\nu(s) =$

$\left(\int_{-1}^{M(s)} tK(t)dt\right)^2$, where $M(s) = \min(s,1)$. The MSE of the cut and normalize estimator, $f^{cn}(x;h)$, is

$$\mathrm{MSE}_x(f^{cn},f) = \frac{f(x)}{nh}\mu(x/h) + h^2\nu(x/h)f'(x)^2,$$

and the MISE is

$$\begin{aligned}
\mathrm{MISE}(f^{cn},f) &= \frac{1}{nh}\int_0^1 \mu(x/h)f(x)dx + h^2\int_0^1 \nu(x/h)f'(x)^2dx \\
&= \frac{\mu(1)}{nh}\int_h^1 f(x)dx + \frac{1}{nh}\int_0^h \mu(x/h)f(x)dx + h^2\int_0^h \nu(x/h)f'(x)^2dx \\
&= \frac{\mu(1)}{nh} + O(n^{-1}) + \frac{1}{n}\int_0^1 \mu(y)f(hy)dy + h^3\int_0^1 \nu(y)f'(hy)^2dy \\
&= \frac{\mu(1)}{nh} + \frac{1}{n}\int_0^1 \mu(y)\left(f(0) + R_1(hy)\right)dy \\
&\qquad + h^3\int_0^1 \nu(y)\left(f'(0)^2 + R_2(hy)\right)dy + O(n^{-1}) \\
&= \frac{\mu(1)}{nh} + h^3f'(0)^2\int_0^1 \nu(y)dy + O(n^{-1}) + o(h^3),
\end{aligned}$$

where $R_1(hy) = o(1)$ and $R_2(hy) = o(1)$. Note that, $\forall\epsilon > 0$, $|R_1(hy)| < \epsilon$ if $|hy| < \delta$ which will occur if $|h| < \delta$ since $0 \le y \le 1$. Thus $|\int_0^1 \mu(y)R_1(hy)dy| \le \int_0^1 |\mu(y)R_1(hy)|dy \le S\int_0^1 |R_1(hy)|dy \le S\int_0^1 \epsilon dy \le S\epsilon$ if $|h| < \delta$, where $S = \sup_{0\le t\le 1}|\mu(t)|$. Thus $\int_0^1 \mu(y)R_1(hy)dy = o(1)$ as $h \to 0$. The best rate of convergence of MISE is $O(n^{-3/4})$, not $O(n^{-4/5})$ as normally expected for a kernel of characteristic exponent $r = 2$. The poor behavior of MSE at the ends dominates the entire MISE calculation. Such results call the use of the cut and normalize kernel into question. It should be noted that this result is in contradiction to the statement of Gasser and Müller (1979); "...end effects may dominate the global asymptotic behavior (for nonparametric regression). Note that this problem does not arise for kernel estimation of densities."

Another method of dealing with the boundary effect is the method of reflection, which is detailed by Schuster (1985). He defines new random variables, $Y_i = S_iX_i$, where $P[S_i = 1] = P[S_i = -1] = \frac{1}{2}$ and the S_i's are independent of the X_i's. The density function of Y is $f_Y(y) = f(|y|)/2$ which is continuous

at $y = 0$ and so kernel methods should be satisfactory. If K is symmetric, the estimate of f does not depend on the S_i's and is

$$f^r(x;h) = \frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{x-X_i}{h}\right) + K\left(\frac{x+X_i}{h}\right)\right].$$

In terms of the usual kernel density estimation formula (2.4.1), the kernel is found to be

$$K_s^r(t) = \begin{cases} [K(t) + K(2s-t)]I_{[-1,s]}(t), & \text{for } 0 \le s \le 1; \\ K(t), & \text{for } s \ge 1, \end{cases}$$

where $s = x/h$ as before. Referring to Figure 3, this method amounts to "folding over" the unused portion of the kernel in region (a) back into the region where the data lies.

Given a standard kernel representation for $f^r(x;h)$, the MSE and MISE can be examined. It can be verified that $\int_{-1}^{1} K_s^r(t)dt = 1$ for all s but that $\int_{-1}^{1} tK_s^r(t)dt \neq 0$ for $0 \le s < 1$. Again, one expects to observe the degraded MISE characteristics the cut and normalize method experiences. Schuster does point out, however, that f^r is non-negative, integrates to 1 and is asymptotically normal.

Both the cut and normalize kernel and the reflection kernel are boundary kernels. That is, they are kernels which change their shape when estimating f near the boundary. One wonders whether it is possible to define a boundary kernel in such a way that the normal MSE and MISE properties are preserved. At least two authors have investigated this possibility. Rice (1984) states the problem in terms of nonparametric regression but notes that the results translate directly to the density estimation problem. In this discussion, his results shall be given in terms of density estimation. Since the cause of difficulty is the bias term, Rice uses a jackknife approach to reduce the bias. This is similar in spirit to the approach Schucany and Sommers (1977) take to reduce the bias of kernel estimators in the non-boundary case. If f admits a second order Taylor's series expansion, $\mathrm{E}[f^k(x;h)]$ can be represented as

$$\mathrm{E}[f^k(x;h)] = \frac{1}{h}\int_0^1 K\left(\frac{x-u}{h}\right) f(u)du,$$

$$= \int_{(x-1)/h}^{x/h} K(y) f(x - hy) dy.$$

Now suppose $0 \leq x \leq h$, so that

$$\mathrm{E}[f^k(x;h)] = \int_{-1}^{s} K(y)[f(x) - f'(x)hy + \frac{1}{2}f''(x)h^2y^2 + o(h^2y^2)]dy,$$

where $s = x/h$, thus

(2.4.4) $$\mathrm{E}[f^k(x;h)] = f(x)w_0(s) - hf'(x)w_1(x) + \frac{1}{2}h^2f''(x)w_2(x) + o(h^2),$$

where $w_i(s) = \int_{-1}^{s} t^i K(t)dt$ for $i = 0, 1, 2$. Rice (1984) suggests defining a new estimator by

$$f^j(x;h,\alpha,\beta,a) = \frac{1}{a}\Big(f^k(x;h) - \beta[f^k(x;h) - f^k(x;\alpha h)]\Big).$$

Using (2.4.4), one finds the expected value of f^j to be:

$$\begin{aligned}\mathrm{E}[f^j(x;h,\alpha,\beta,a) = \frac{1}{a}\Big(&[(1-\beta)w_0(s) + \beta w_0(s/\alpha)]f(x) \\ &- [(1-\beta)w_1(s) + \alpha\beta w_1(s/\alpha)]hf'(x) \\ &+ \frac{1}{2}[(1-\beta)w_2(s) + \alpha^2\beta w_2(s/\alpha)]h^2f''(x)\Big) + o(h^2).\end{aligned}$$

The parameters α, β, and a need to be chosen so that this expectation is $f(x) + const.h^2$. This is accomplished by setting

$$a = (1-\beta)w_0(s) + \beta w_0(s/\alpha), \qquad \beta = \frac{w_1(s)}{w_1(s) - \alpha w_1(s/\alpha)}.$$

The parameter α is still free; Rice suggests setting $\alpha = 2 - s$ so that one always smooths over an interval of length $2h$. The kernel which defines f^j is

$$K_s^j(t) = \begin{cases} [(1-\beta_s)K^a(t) + (\beta_s/\alpha_s)K^b(t)]/a_s, & \text{for } 0 \leq s \leq 1; \\ K(t), & \text{for } s \geq 1, \end{cases}$$

where

$$K^a(t) = K(t)I_{[-1,s]}(t) \text{ and } K^b(t) = K(t)I_{[-\alpha,s]}(t).$$

It isn't hard to show that K_s^j satisfies $\int_{-1}^{1} K_s^j(t)dt = 1$ and $\int_{-1}^{1} tK_s^j(t)dt = 0$ for all s.

Figure 5 presents K_s^j for s =0, 0.25, 0.5, and 0.75 for $K(t)$ equal to the biweight kernel. Notice that these kernels do eventually have negative regions as

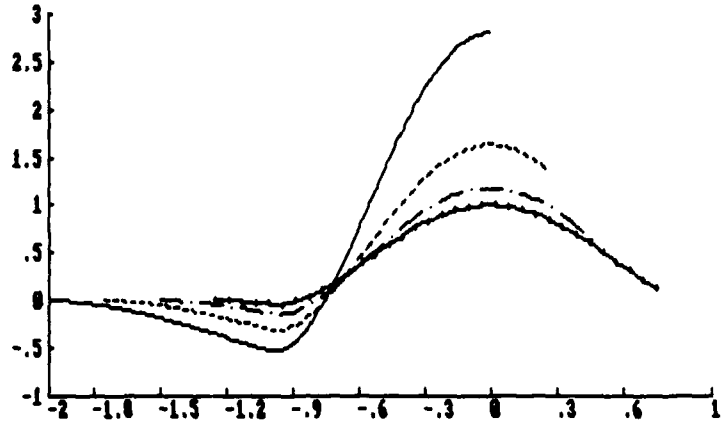

Fig. 5. *Rice's boundary kernels for s=0,* 0.25, 0.5, *and* 0.75. *The solid line is* $s = 0$*; the broken line,* $s = 0.25$*; the broken line with points interspersed,* $s = .5$*; and the solid line with nodes,* $s = .75$.

s decreases. This is the price that must be paid to keep the first moment equal to zero throughout. The usual MSE and MISE results will apply to this kernel. It appears that one has the choice of non-negative estimates with bias of order $O(h)$ or potentially negative estimates with bias of order $O(h^2)$.

From (2.4.4) it can be seen that the essential conditions for bias reduction are

$$\text{(2.4.5)} \qquad \begin{aligned} &\text{i.} \quad \int_{-1}^{s} K_s(t)dt = 1, \\ &\text{ii.} \quad \int_{-1}^{s} tK_s(t)dt = 0, \end{aligned}$$

for all $0 \leq s \leq 1$. Approaching the problem from this standpoint, Gasser and Müller (1979) propose a boundary kernel of the form

$$K_s^{gm}(t) = \begin{cases} (\theta_s + \phi_s t)K(t)I_{[-1,s]}(t), & \text{for } 0 \leq s \leq 1; \\ K(t), & \text{for } s \geq 1, \end{cases}$$

where θ_s and ϕ_s are chosen to be continuous functions of f such that constraints (2.4.5) hold for all s and $\theta_1 = 1$ and $\phi_1 = 0$. Using the biweight kernel and substituting the form of K_s^{gm} into the constraints (2.4.5), θ_s and ϕ_s are the

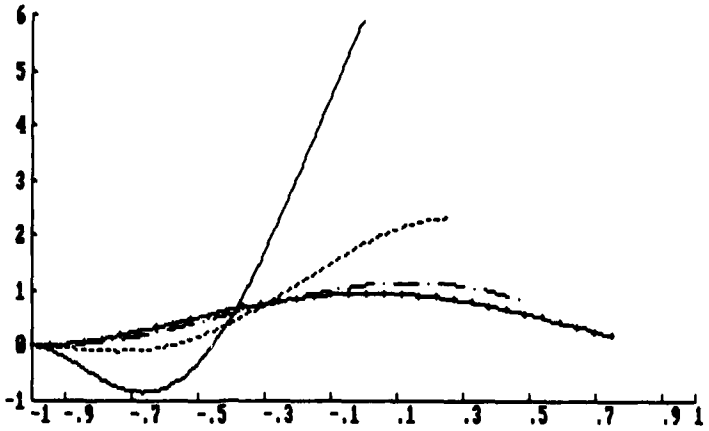

Fig. 6. *Gasser and Müller's boundary kernels for s=0,* 0.25, 0.5, *and* 0.75. *The solid line is* $s = 0$*; the broken line,* $s = 0.25$*; the broken line with points interspersed,* $s = 0.5$*; and the solid line with nodes,* $s = 0.75$.

solutions to the following linear equations:

$$\begin{aligned}
&\left[\frac{1}{5}(s^5+1) - \frac{2}{3}(s^3+1) + s + 1\right]\theta_s \\
&\qquad + \left[\frac{1}{6}(s^6-1) - \frac{1}{2}(s^4-1) + \frac{1}{2}(s^2-1)\right]\phi_s = 1 \\
&\left[\frac{1}{6}(s^6-1) - \frac{1}{2}(s^4-1) + \frac{1}{2}(s^2-1)\right]\theta_s \\
&\qquad + \left[\frac{1}{7}(s^7+1) - \frac{2}{5}(s^5+1) + \frac{1}{3}(s^3+1)\right]\phi_s = 0.
\end{aligned} \tag{2.4.6}$$

Graphs of $K_s^{gm}(t)$ for s=0, 0.25, 0.5, and 0.75 are displayed in Figure 6. Again notice that the kernel must eventually have negative regions to satisfy the constraints (2.4.5). Although for given s the support of K_s^{gm} and K^j are different, their shapes are similar in that one can see a progressive and continuous deformation of the original kernel. Despite needing to solve (2.4.6) for θ_s and ϕ_s, the Gasser-Müller kernel may be somewhat easier to work with since its support depends on s on only one side. Both kernels are rational functions of s.

The Gasser-Müller boundary kernel for the right hand endpoint, $x = 1$, is easily obtained. One could derive the expression for the expected value of the estimator for the right endpoint. One would then notice the same problem except

that the integrals are from $-s$ to 1 instead of -1 to s. In this case, s is given by $s = (1-x)/h$. Due to the symmetry of the problem, it is easily shown that $K^r_{s'}(w) = K^l_s(-w)$, where $s' = (1-x)/h$, $s = x/h$, K^l is the left hand boundary kernel and K^r is the right hand boundary kernel.

2.4.4. *Orthogonal Series Methods.* Orthogonal series methods are reviewed in this subsection. Many orthogonal series lend themselves naturally to the estimation of densities on $[0,1]$. Care must be given to the choice of orthogonal functions, however, since this choice can have profound effects upon the properties of the estimator. Although orthogonal series ideas are taken up in Section 3, the traditional methods are seen in this subsection not to be exactly what is sought.

The expansion of non-random functions by orthogonal series is a commonly used technique of mathematical analysis [see Stromberg (1981)]. Cencov (1962) was the first to suggest that such techniques could be useful in the area of density estimation.

Start by assuming that f has support $[0,1]$. There do exist orthonormal bases on $\mathbb{R}$, such as the Hermite functions, however, these are not of primary interest here. Let $\{\phi_j(x)\}_{j=1}^{\infty}$ be a complete orthonormal basis for $\mathcal{L}_2[0,1]$, which is the space of all functions on $[0,1]$ which are square integrable with respect to the weight function, $w(x)$. A basis is orthonormal with respect to the weight function $w(x)$ if

$$\int_0^1 \phi_j(x)\phi_k(x)w(x)dx = I(j=k),$$

where $I(j=k)$ is the (Kronecker's delta) indicator function. The basis is complete, if for all $g \in \mathcal{L}_2[0,1]$, there exists a sequence of constants $\{a_k\}_{k=1}^{\infty}$ such that

$$\|g - \sum_{k=1}^{n} a_k\phi_k\| \to 0 \text{ as } n \to \infty.$$

The sequence $\{a_k\}$ is given by $a_k = \int_0^1 \phi_k(x)g(x)w(x)dx$, and $\|\cdot\|$ is the $\mathcal{L}_2[0,1]$ norm with respect to the weight function, $\|g\|^2 = \int_0^1 g(x)^2 w(x)dx$.

Assume that the density f admits the series representation,

$$f(x) \sim \sum_{k=1}^{\infty} a_k \phi_k(x),$$

where $a_k = \int \phi_k(x) w(x) f(x) dx$. The orthogonal series estimate of the density is defined to be

$$f^o(x;\lambda) = \sum_{k=1}^{\infty} \lambda_k \hat{a}_k \phi_k(x), \tag{2.4.7}$$

where

$$\begin{aligned} \hat{a}_k &= \int_0^1 \phi_k(x) w(x) dF_n(x) \\ &= \frac{1}{n} \sum_{i=1}^{n} w(X_i) \phi_k(X_i). \end{aligned}$$

The sequence $\{\lambda_k\}$ is the smoothing parameter, which one expects to decline to 0 as k becomes large. Several forms for the sequence have been suggested. For the moment, it will be assumed that

$$\lambda_k = \begin{cases} 1, & \text{if } k \leq m; \\ 0, & \text{otherwise.} \end{cases} \tag{2.4.8}$$

The estimator of equation (2.4.7) now becomes

$$f^o(x;m) = \sum_{k=1}^{m} \hat{a}_k \phi_k(x). \tag{2.4.9}$$

The parameter m is referred to as the truncation point of the series and plays the role of smoothing parameter.

Kronmal and Tarter (1968) consider estimators of form (2.4.9). Although any orthonormal basis $\{\phi_k\}$ will do, they focus on trigonometric (or Fourier) series. There are at least four distinct bases involving trigonometric functions; these are

i. $\phi_k(x) = \cos \pi k x, \ k \geq 0;$
ii. $\phi_k(x) = \sin \pi k x, \ k \geq 1;$
iii. $\phi_k(x) = (\cos \pi k x, \sin \pi k x), \ k \geq 1;$
iv. $\phi_k(x) = (\cos 2\pi k x, \sin 2\pi k x), \ k \geq 0.$

These will be referred to as bases (i), (ii), (iii), and (iv), respectively.

Kronmal and Tarter (1968) give results for (i), (ii), and (iii), whereas (iv) is recognized as the basis function used by Parzen (1983) in the AR spectral approach to estimating $d_\lambda(u)$ discussed in Subsection 2.3.4 and by Parzen (1979) and Carmichael (1984) in a general density estimation setting. Trigonometric functions are often used as they are convenient, easy to calculate and their properties are well known. Kronmal and Tarter also note that since trigonometric functions differentiate and integrate to other trigonometric functions, one does not have to choose between an orthogonal series expansion of F and f. For the trigonometric bases, the weight function is $w(x) = 2$. Other orthonormal bases do exist for $\mathcal{L}_2[0,1]$; the Legendre polynomials are an example.

Although bases (i) through (iv) are complete in $\mathcal{L}_2$, it is well known [see Wahba (1975) or Hall(1981)] that the estimates obey certain conditions. Though the details are different, it was seen in Subsection 2.3.4 that basis (iv) imposes $d_\lambda(0) = d_\lambda(1)$ upon the estimates, as it does on f^o. Basis (ii) also imposes this condition. In the case that $f(0) \neq f(1)$, these trigonometric series exhibit Gibb's *phenomenon*. The estimates will tend to be very wiggly near $x = 0, 1$, and in fact f^o will be estimating $(f(0) + f(1))/2$ at the points $x = 0, 1$. This poses no difficulty to the $\mathcal{L}_2$ convergence of the series since the $\mathcal{L}_2$ norm is insensitive to pointwise errors. Hall (1981) discusses Gibb's phenomenon. He finds that the rate of convergence of MISE can be as bad as $O(1/\sqrt{n})$. Newton (1988), page 77, has an excellent general discussion of Gibb's phenomenon. The problem arises in part due to the choice of λ_k as (2.4.8). In time series analysis, it is usual to give the λ_k's a damped form to reduce this problem. Basis (i) imposes the following end conditions on the derivatives, $f^{o(2k-1)}(0; m) = f^{o(2k-1)}(1)$, for $k \geq 1$. Hall (1983b) finds that this series is far more resistant to Gibb's phenomenon than any of the other three.

For the cosine series, basis (i), the density estimator is found by applying equation (2.4.9), with the exception that the weight function $w(x) = 1$ not $w(x) = 2$ is appropriate for $k = 0$. Bearing this in mind, (2.4.9) becomes

$$f^i(x; m) = \frac{a_0}{2} + \sum_{k=1}^{m} \hat{a}_k \cos \pi k x,$$

where

$$\hat{a}_k = \frac{2}{n}\sum_{i=1}^{n} \cos \pi k X_i \text{ for } k \geq 0.$$

In the case of bases (iii) and (iv), the terms are taken in sine/cosine pairs so that the estimator is

$$f^{iii,iv}(x;m) = \frac{a_0}{2} + \sum_{k=1}^{m} \hat{a}_k \phi_k^c(x) + \sum_{k=1}^{m} \hat{b}_k \phi_k^s(x),$$

where

$$\hat{a}_k = \frac{2}{n}\sum_{i=1}^{n} \phi_k^c(X_i) \text{ for } k \geq 0,$$

$$\hat{b}_k = \frac{2}{n}\sum_{i=1}^{n} \phi_k^s(X_i) \text{ for } k \geq 1.$$

For basis (iii), $\phi_k^c(x) = \cos \pi k x$, $\phi_k^s(x) = \sin \pi k x$, and $a_0 = 0$; for basis (iv), $\phi_k^c(x) = \cos 2\pi k x$, $\phi_k^s(x) = \sin 2\pi k x$, and $a_0 = 2$.

One cannot be assured that f^o constructed from any of these bases will be a valid density. In particular, f^o can become negative. Kronmal and Tarter (1968) make two points concerning this issue. First they note that in all their simulations they did not come up with a negative estimate. This point seems somewhat weak as it is based solely on a handful of simulated data sets. Their second point is that negative estimates are not a complete anathema. Negative estimates should serve as a warning that inference in the negative region is hazardous; that there is insufficient data in the region. This second point seems a much more appealing response. The estimate, $f^o(x;m)$, at a point is a random variable taking on values in $\mathbb{R}$; if f is small at x, there is no reason to be surprised that $f^o(x;m)$ should be negative.

Notice that f^i and f^{iv} do integrate to 1. If the estimate is non-negative, then the result is a density. If the estimate has negative regions, than the fact that it integrates to 1 is of no interest. The key condition is non-negativity: given a non-negative (and integrable) estimate one can always normalize it to arrive at a probability density.

As with histograms and kernel density estimators, the performance of $f^o(x;m)$ is usually measured by MSE and MISE. Kronmal and Tarter give the

MISE as

$$\text{MISE}(f^o, f) = \sum_{k=1}^{m} \text{Var}(\hat{a}_k) + \sum_{k=1}^{\infty} a_k^2.$$

Kronmal and Tarter (1968) establish that if $m = o(\sqrt{n})$, then $\text{MISE}(f^i, f) \to 0$ as $n \to \infty$. Using basis (iv), Hall (1981) derives the following relations for MISE assuming that f possesses r derivatives and that $f^{(j)}(0) = f^{(j)}(1)$, for $j = 0, \ldots, r-1$,

$$\text{MISE}(f^{(iv)}, f) = \frac{m}{n\pi} + \frac{1}{(2r+1)\pi}[f^{(r)}(0) - f^{(r)}(1)]m^{-2r-1} + o(mn^{-1} + m^{-2r-1}).$$

The best rate of convergence of MISE is $O(n^{(2r+1)/(2r+2)})$ which is $O(n^{5/6})$ for $r = 2$. This rate improves somewhat that of the boundary kernel (Gasser-Müller or Rice), which is $O(n^{4/5})$, for $r = 2$. The improved rate, however, is obtained only at the cost of requiring f and its first $r-1$ derivatives to be periodic.

Hall derives a similar result for basis (ii), by requiring f to possess $2r$ derivatives satisfying $f^{(2j)}(0) = f^{(2j)}(1) = 0$, for $j = 0, \ldots, r$. A rate of $O(n^{(4r+1)/(4r+2)})$ is then obtained. For basis (i), he requires f to possess $2r+1$ derivatives which satisfy, $f^{(2j+1)}(0) = f^{(2j+1)}(1) = 0$, for $j = 1, \ldots, r$. In this case, the best rate of convergence of MISE is $O(n^{(4r+3)/(4r+4)})$. Notice that in this case the restrictions apply only to the derivatives of f, not to the end values of f itself. This result must be related to the observation that series (i) is the most resistant to Gibb's phenomenon.

Return now to the general definition of the orthogonal series estimator given by (2.4.7). Until now, the special form of λ_k of equation (2.4.8) has been assumed. One wonders if there might be a better choice of weights. Watson (1969) finds the weights which minimize the MISE of f^o to be

$$\lambda_k = \frac{a_k^2 / \text{E}\left[\phi_k(X)^2\right]}{\frac{1}{n}\left(1 + (n-1)a_k^2 / \text{E}\left[\phi_k(X)^2\right]\right)}.$$

He notes that for fixed k that $\lambda_k \to 1$ as $n \to \infty$. He concludes that ordinary truncation will probably be sufficient if the a_k's are large compared to $\text{Var}(\phi_k(X))/n$ for $k \leq m$ and negligible for $k > m$.

Wahba (1981) defines an estimator similar in spirit to Watson's using basis (iv). She assumes the same periodicity conditions on f and its first $m-1$ derivatives as Hall (1981), above. She defines the weights, λ_k, parametrically and gives them a Bayesian interpretation. The final form of her estimator is

$$f_{n,\lambda,m}(x) = 1 + 2\sum_{k=1}^{n/2} \frac{1}{1+\lambda(2\pi k)^{2m}} \left(\hat{a}_k \cos 2\pi kx + \hat{b}_k \sin 2\pi kx\right).$$

Wahba shows that if, $\lambda = an^{-2m/(2m+1)}$, for some constant, a, that MISE = $O(n^{-2m/(2m+1)})$.

As a final point on orthogonal series density estimation, it seems natural to ask if it bears any relation to kernel density estimation. The answer is yes and in the case of basis (iii) Kronmal and Tarter (1968) give the relation as

$$f^{(iii)}(x;m) = \frac{1}{n}\sum_{i=1}^{n} \delta_m(X_i - x),$$

where

$$\delta_m(x) = \frac{\sin[(2m+1)\pi x/2]}{\sin[\pi x/2]},$$

is known as the Dirichlet kernel. There is no explicit bandwidth for the Dirichlet kernel; instead m plays the role of the smoothing parameter. The relation between m and h, the bandwidth of the usual kernel estimator, is approximately $h \sim 1/m$. Figure 7 displays $\delta_m(x)$ for $m = 2, 4, 8$ and 16. Graphically, it is easy to see the role of m. Note that this kernel is not unimodal and not non-negative. Interestingly, the usual kernel approaches a delta function (unbounded at zero, zero elsewhere) as $h \to 0$, whereas δ_m does not as $m \to \infty$. Although δ_m becomes unbounded at zero, the side lobes never decay to zero. This raises an interesting question: Should one choose an orthogonal series that's convenient and not worry about the kernel representation or conversely? It is hard to imagine anyone approaching this problem from the kernel perspective actually choosing to use the Dirichlet kernel.

2.4.5. *AR/ARMA Methods.* The last categories of density estimates to be examined are the autoregressive (AR) and autoregressive moving average

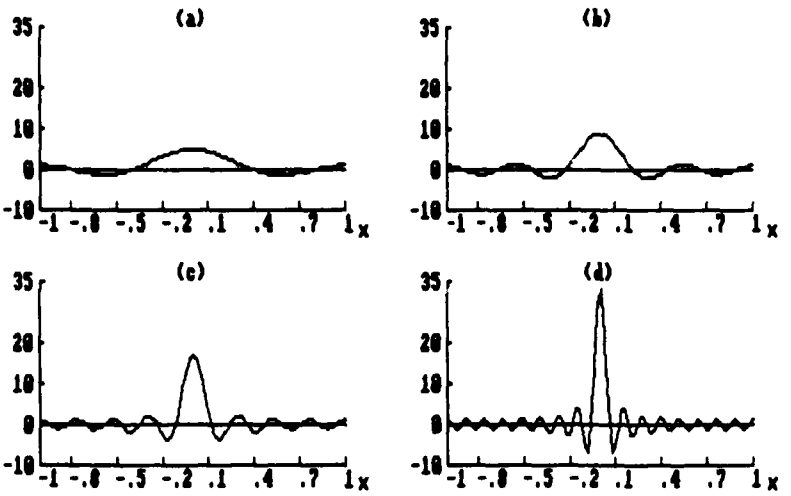

Fig. 7. *The Dirichlet kernel. Figure* (a) *is constructed with* $m = 2$*; Figure* (b)*,* $m = 4$*; Figure* (c)*,* $m = 8$*; and Figure* (d)*,* $m = 16$*.*

(ARMA) estimators. The AR and ARMA methods are natural for densities of compact support. It is seen in this subsection that the AR approach imposes certain restrictions on the estimated density. The ARMA approach is somewhat less restrictive and contains as a special case the cosine based Fourier series estimate.

The form of the AR estimator has been given in Subsection 2.3.4, however, it is repeated here for the sake of completeness. The estimator, $f^{AR}(x;m)$, is defined by

$$f^{AR}(x;m) = \sigma_m^2 \left| 1 + \sum_{j=1}^{m} \alpha_j e^{2\pi i j x} \right|^{-2},$$

where the α_j's are complex-valued and $|\cdot|^2$ denotes the complex squared modulus. The α_j's and σ_m^2 are estimated from the data. The discussion of Subsection 2.3.4 carries over exactly with the exception of the estimates of the pseudo-correlations. In this framework, the pseudo-correlations are estimated by

$$\hat{\rho}_k = \int_0^1 e^{2\pi i k x} dF_n(x), \text{ for } k = 0, 1, 2, \ldots, .$$

The only difference is that dF_n replaces dD_N in the definition of $\hat{\rho}_k$ in Subsection 2.3.4. Carmichael (1984) obtains a consistency result for $f^{AR}(x;m)$ by allowing

m to grow with n at an appropriate rate. The estimate, $f^{AR}(x;m)$, is itself a density; it is non-negative and integrates to one. As noted in Subsections 2.3.4 and 2.4.4, the estimates have the property that $f^{AR}(0;m) = f^{AR}(1;m)$. One should expect the estimates to exhibit bias if this condition is not met by the underlying density.

Hart (1988) suggests what he calls an ARMA density estimate. It is defined as

$$f^{ARMA}(x;m,\alpha) = (\hat{\beta}_0 + 2\sum_{j=1}^{m} \hat{\beta}_j \cos \pi j x)|1 + \alpha e^{\pi i j x}|^{-2}$$
$$= f^{i}(x;m) + 2\text{Real}\Big([\hat{a}_m \alpha e^{i(m+1)x}]/[1 - \alpha e^{i\pi x}]\Big),$$

where $\hat{a}_m$ is the estimate of the m^{th} Fourier coefficient of f and f^{i} is defined in Subsection 2.4.4. f^{ARMA} is called an ARMA density estimate because the form of f^{ARMA} is similar to that of an ARMA(1,m) spectral density. Hart specifically uses a cosine based series to minimize the Gibb's phenomenon experienced by the estimate. The pair, (m,α), constitute the smoothing parameter. Since $f^{ARMA}(x;m,0) = f^{i}(x;m)$, f^{ARMA} is a more general estimator than the cosine Fourier series.

Hart derives exact and approximate MISE results. In particular, if the Fourier coefficients, $\{a_k\}$, of f and $\alpha = \alpha(m)$, obey either (a) $j^{\rho}a_j \to K \neq 0$ as $j \to \infty$ and $m(1-\alpha) \to c > 0$ as $m \to \infty$ or (b) $(-1)^j j^{\rho} a_j \to K \neq 0$ as $j \to \infty$ and $m(1-\alpha) \to c > 0$ as $m \to \infty$, for $\rho > \frac{1}{2}$, the best rate of convergence of MISE is $O(n^{1/2\rho-1})$ for both f^{ARMA} and f^{i}. In the case where $c = \rho$, Hart shows that $\text{MISE}(f^{ARMA}, f)/\text{MISE}(f^{i}, f) < 1$; that is, even though the two obtain the same rate, the constant in $O(n^{1/2\rho-1})$ for f^{ARMA} is smaller. This result is reasonable given the class of cosine Fourier estimates is a subset of the ARMA estimates.

2.4.6. *Choosing the Smoothing Parameter.* Each estimator discussed in Subsections 2.4.2 through 2.4.5 is indexed by some sort of smoothing parameter. Let s denote a generic smoothing parameter. In this subsection, various methods of choosing s will be discussed.

One of the first methods suggested is to choose s to minimize MISE. Unfortunately, in each case the optimal value depends in some way on the unknown density, f. For the histogram, one needs $\int f'(x)^2 dx$; for kernels, $\int f''(x)^2 dx$ (assuming $r = 2$); for orthogonal series, the Fourier coefficients. Several suggestions have been made to overcome this difficulty. One method estimates the unknown quantity by the value it would have if f falls in some parametric class; see Scott (1979) and Silverman (1986). In the case of kernel estimates, one can estimate the unknowns from the data nonparametrically; see Woodroofe (1970) and Scott, Tapia and Thompson (1977). The parametric methods generally perform adequately if f resembles the assumed family (for example, if f is unimodal). The nonparametric methods fail to perform well in the simulation studies Bowman (1985) conducts.

The second major class of selection methodologies could be termed selection through optimization. In this technique, s is chosen as the optimizing value of some objective function. There are two general types of objective functions; the likelihood function and estimates of MISE. Duin (1976) introduces the first, which is termed likelihood cross-validation. The objective function is defined as

$$L(s) = \prod_{i=1}^{n} \hat{f}_{(i)}(X_i; s),$$

where $\hat{f}_{(i)}$ is the estimate of f calculated with the i^{th} observation omitted. The parameter s is chosen to maximize $L(s)$. The usual likelihood function, $\prod \hat{f}(X_i; s)$, is not employed since it typically leads to a degenerate choice of s (*i.e.* $s = 0$ or ∞). In the case of kernel estimators, Chow, Geman and Wu (1983) prove that if f is bounded and of compact support and h is chosen to maximize $L(h)$, then $\mathrm{ISE}(f^k, f) \to 0$ a.s. as $n \to \infty$. Here, ISE is the integrated squared error, $\mathrm{ISE}(f^k, f) = \int [f^k(x; h) - f(x)]^2 dx$. In general circumstances, the restrictions on f are of concern, although not so here.

In the second method, an estimate of MISE is minimized with respect to s. For trigonometric series, it is possible to estimate MISE, or its increments, directly; see Kronmal and Tarter (1968), Tarter and Kronmal (1976), Hart (1985), Diggle and Hall (1986), and Wahba (1981). Rudemo (1982) and Bowman (1984)

introduce least squares cross-validation (LSCV), which has application to a wide range of estimators. The objective function is given by

$$\mathrm{LSCV}(s) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{(i)}(x;s)^2 dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{(i)}(X_i;s).$$

Rudemo (1982) shows that LSCV is an unbiased estimator of $\mathrm{MISE}(\hat{f},f) - \int f(x)^2 dx$. Since the last term does not depend on s, it is hoped that minimizing LSCV with respect to s will be like minimizing MISE. Hall (1983a) and Stone (1984) give results for kernel estimates concerning the behavior of f^k when h is so chosen. In particular, assuming only that f is bounded, Stone shows that

$$\frac{\int [f^k(x;\hat{h}) - f(x)]^2 dx}{\int [f^k(x;h^\dagger) - f(x)]^2 dx} \to 1 \text{ a.s. as } n \to \infty,$$

where $\hat{h}$ is the minimizer of LSCV and $h^\dagger$ minimizes ISE.

Of all the methods discussed, LSCV is probably the most widely used, although one should not regard LSCV as a panacea. Silverman (1986), page 51, points out that for kernel estimation LSCV can lead to a degenerate choice of h if the observations are discretized. It is also well recognized [see Hart (1988,1985) and Scott and Terrell (1987)] that LSCV tends to substantially undersmooth about 5% to 20% of the time. Nonetheless, least squares cross-validation is a useful and general tool.

2.4.7. *Choice of Estimator.* The estimator to be used in this dissertation is the boundary kernel of Gasser and Müller. The rationale for choosing a kernel based estimator and specifically the Gasser-Müller boundary kernel is detailed below.

The histogram is not used because it is felt that it does not convey information well. It is inherently rough and discrete, yet it is estimating an object which is continuous and smooth. A smooth estimator is desired. Further, the rate of convergence of MISE falls well below that of other techniques examined.

A trigonometric series is not employed because one is led to kernel representations for the estimate that use poor kernels. To answer the question posed at the end of Subsection 2.4.4, it is better to choose what seems an appropriate

kernel; one that leads to sensible estimates and has sensible properties. This is not to say that orthogonal series have been abandoned altogether. In Section 3, a series representation for the Gasser-Müller boundary kernel is obtained. Rather, it is to say that it is better to let the kernel determine the orthogonal series. Kernels seem to be more easily examined as to their implications for the estimate. Further, the kernel based orthogonal series is indexed by the bandwidth, h. In Section 3, where the orthogonal series will be used to construct score functions for linear rank statistics, this will be very convenient. It will be seen that the tail behavior of the score functions varies with h so that distinct values of h correspond to optimal scores for different distributions. If an ordinary orthogonal series were used, it would be necessary to change the series to achieve such an effect. Since the ARMA method is very nearly an orthogonal series method much the same reasoning applies.

The cut and normalize and reflection boundary kernels are not used because of their greater bias. If is felt that it is worth trading bias reduction for the guarantee of non-negative estimates. The discussion of the interpretation of negative estimates in Subsection 2.4.4 removes some of the onus of the situation. Further, one needs to examine the potential uses of an estimate of $d_\lambda(u)$. Even though $d_\lambda(u)$ is a density, it won't be used for simulation, nor will probabilities be calculated. Recalling $d_\lambda(u)$'s interpretation as a likelihood ratio, the important feature of $d_\lambda(u)$ is its shape–which regions are large relative to others and relative to 1. Regions which are negative are to be interpreted as having little or no content. In these circumstances, there is far less need to require the estimate to itself be a density function. Rather, it is preferable to have an improved estimate in terms of MISE. Finally, there is not a lot to choose between Gasser and Müller's boundary kernel and that of Rice. Both have the same asymptotic representation and broadly similar shapes. It is somewhat more compact to write down the Gasser-Müller boundary kernel and so it is selected.

3. ESTIMATION AND TESTING

3.1. Introduction

3.1.1. *Introduction.* Results for the boundary kernel estimation of the comparison density and tests of its uniformity are presented in this section. Subsection 3.2 describes results concerning the estimation phase of the process. Subsection 3.2.2 gives more traditional results for the estimator, including its asymptotic normality under H_o and its consistency under general alternatives. Both these results are derived under a shrinking bandwidth. The invariance properties of the estimator are also detailed.

Subsection 3.2.3 defines a stochastic process based on the boundary kernel estimate. This stochastic process is called the kernel density process. Under a fixed bandwidth, the weak convergence of this process to a limiting Gaussian process is proved. A convenient representation for the limiting process is given. Properties of these processes are explored. They are seen to be continuous with probability 1. The null covariance kernel, which is the covariance kernel of the limiting process under H_o, is given. The quality of the approximation of the limiting distribution under fixed h is compared to that of shrinking h under H_o. A simulation study is conducted to carry out this comparison. The results indicate that the fixed h approximation is superior.

Subsection 3.3 gives results concerning the testing phase of the procedure. During the study another estimator of the comparison density suggests itself. Subsection 3.3.2 details the statistic $\hat{\varphi}^2_{N,h}$ which is the square of the $\mathcal{L}_2$ norm between the boundary kernel estimator and 1. Although it is a statistic and so is not what is sought, an analysis of its distribution leads to the idea of the components of the kernel density process. These components are similar in spirit to those of the Cramér-von Mises and Anderson-Darling statistics described in Subsection 2.2.5. Subsection 3.3 investigates the components in depth. Also investigated are those concepts required to define the components such as the null covariance kernel and its eigenvalues and eigenfunctions. A numerical procedure

to estimate the eigenfunctions and eigenvalues is given and a check of its accuracy performed. The properties of the components are worked out. They are seen to be generalized Fourier coefficients and linear rank statistics. The joint convergence in distribution of the sample components to their limiting counterparts is proved. A small sample correction to the means of the components is given. It is seen that the space spanned by the eigenfunctions is of interest. Finally, it is argued that a test based on the first M components which gives equal weight to each will yield more fruitful results than the traditional statistics which are weighted infinite sums of the components.

Subsection 3.3.4 investigates whether the space spanned by the eigenfunctions contains the space in which the kernel density process resides. This condition is seen to be related to the positive definiteness of the null covariance kernel over this space. An equivalent condition which is a Fredholm integral equation of the first kind is given. Unfortunately, it is not possible to check either of these conditions: the equations are too complex. The implications of the eigenfunctions spanning or not spanning the appropriate space are detailed.

Subsection 3.3.5 introduces a new framework for testing the components. From this framework is suggested what is called the subset chi-square test. The traditional tests–the chi-square test and the independent tests method (*i.e.* testing each component independently)–fit into this framework as well. The three tests are compared. The subset chi-square is shown to be a compromise between the other two. It is also seen to possess other desirable properties. It considers the components as groups not just singly. In the case of rejection, it indicates which components are significant. Finally, it lends itself well to graphical display.

Subsection 3.3.6 applies the subset chi-square test to the components. The test suggests an orthogonal series estimate of the comparison density. This orthogonal series estimate is investigated and contrasted to the boundary kernel estimator. The orthogonal series estimate is proved to be a weighted orthonormal series where the weights are the eigenvalues of the null covariance kernel. Subsection 3.3.7 suggests alternate strategies of choosing the bandwidth and truncation point. Also discussed are the pros and cons of automatic selection criteria and

their effect on the testing procedure. Subsection 3.3.8 summarizes the unified procedure.

3.1.2. *Assumptions.* The various assumptions made on the underlying distributions have been scattered throughout the first two sections. They are repeated here for clarity.

1. $X_1, \ldots, X_m$ are iid with distribution function F.
2. $Y_1, \ldots, Y_n$ are iid with distribution function G.
3. The two samples are independent.
4. F and G are absolutely continuous with densities f and g, respectively.
5. The quantiles functions of F and G, Q^F and Q^G, are continuous.
6. If f has support $[a_f, b_f]$ and g has support $[a_g, b_g]$, then f and g are continuous on $(a_f \wedge a_g, b_f \vee b_g)$.
7. Let $\lambda_{(N)} = m/N$, $N = m + n$. Then it is usually assumed that $\lambda_{(N)} \to \lambda_0$ as $m \wedge n \to \infty$ where $0 < \lambda_0 < 1$ but sometimes $\lambda_{(N)} = \lambda_0$ is assumed. It will be pointed out where this latter assumption is used.

3.2. Properties of the Boundary Kernel Estimator

3.2.1. *Introduction.* Subsection 3.2 examines the properties of the boundary kernel estimator of the comparison density function. In Subsection 3.2.2 asymptotic pointwise results are established. The asymptotic normality of the estimator under H_o is established as is its consistency. These results are traditional in the sense that they occur as the bandwidth shrinks to zero at an appropriate rate. The invariance properties of the estimate are also examined. In Subsection 3.2.3, results for the estimator are derived by treating it as a stochastic process on $[0, 1]$. These results are derived under the assumption of fixed bandwidth; that is, the bandwidth does not shrink to zero as the sample sizes increase. A process based on the kernel density estimator called the kernel density process is defined and its weak convergence is shown under this condition. The results of both these subsections are unique because the underlying stochastic process, CD_N, and its limiting process, L, are not those associated with iid random variables for which

density estimation results are usually derived. No such results are known to exist in this framework.

3.2.2. *Pointwise Asymptotic Results.* In this subsection, two main pointwise results are established for the boundary kernel estimate of the comparison density. The first is its asymptotic normality under H_o and the second is its consistency under any alternative. The invariance properties of the estimate are also investigated. Start by defining the boundary kernel estimator of the comparison density as

$$\begin{aligned}\hat{d}_h(w) &= \int_0^1 \frac{1}{h} K_s\left(\frac{w-u}{h}\right) dD_N(u) \\ &= \frac{1}{mh}\sum_{i=1}^{m} K_s\left(\frac{w-(R_i/N)}{h}\right),\end{aligned}$$

where $K_s(w)$ is the Gasser-Müller boundary kernel and $s = s(w,h)$ is given by

$$s(w,h) = \begin{cases} w/h & \text{for } 0 \le w \le h \\ (1-w)/h & \text{for } 1-h \le w \le 1 \\ 1 & \text{otherwise.} \end{cases}$$

The function $s(w,h)$ chooses which of the family of boundary kernels is appropriate. It is understood that when $w < h$ the left hand boundary kernel is selected and when $w > 1-h$ the right hand boundary kernel is selected. The remaining terms have been defined in Subsections 2.2 and 2.3: $D_N(u)$ is the sample comparison distribution function; $N = m+n$ and R_i is the rank of X_i in the pooled sample.

The following theorem concerning the asymptotic normality of $\hat{d}_h(w)$ under H_o can be proved.

Theorem 3.2.1. *If $K(t)$ is a kernel with support $[-1,1]$ satisfying $K(-1) = K(1) = 0$, $K'(-1) = K'(1) = 0$, $K''(-1) = K''(1) = 0$, and $|K''(t)| \le M$ for some $M < \infty$ and $\int_{-1}^{1} K(t)dt = 1$ and if $h \to 0$, $(m \wedge n)h^3 \to \infty$ as $m \wedge n \to \infty$ then under H_o, $\hat{d}_h(w)$ is $\mathrm{AN}(1, \frac{1}{Nh}\frac{1-\lambda_0}{\lambda_0}\int_{-1}^{1} K(t)^2 dt)$ for each $w \in (0,1)$.*

The proof of Theorem 3.2.1 along with the proofs to all the theorems and lemmas stated in Sections 3 and 4 is in Appendix B. The proof of Theorem 3.2.1 is long

and somewhat tedious. Its basic approach is motivated by Chernoff and Savage (1958). The strategy is to write $\sqrt{Nh}(\hat{d}_h(w) - 1)$ as the sum of four terms, two first order terms and two second order terms. The first order terms are shown to converge in distribution to the appropriate random variable while the second order terms are shown to converge in probability to zero.

It should be noted that since the bandwidth, h, is shrinking to zero that boundary effects do not occur asymptotically for each $w \in (0,1)$. This is so since for each fixed w, boundary effects are experienced only if $w < h$ or $w > 1 - h$. Since $h \to 0$, these effects will cease at some point. In the proof of Theorem 3.2.1, one can ignore the Gasser-Müller modification of K and concentrate solely on K itself.

Comparing Theorem 3.2.1 with the types of conditions one normally sees for kernel estimators in the iid case, one notes (a) additional conditions on K and (b) $(m \wedge n)h^3 \to \infty$ instead of $Mh \to \infty$ (for a random sample of size M in the iid case). This comparison is of interest since under H_o one might reasonably expect the normalized ranks, $R_1/N, \ldots, R_m/N$, to behave like a random sample from a uniform, $U(0,1)$, distribution. Indeed, when viewed from a process point of view, they have the same limiting empirical process up to the multiplicative constant $\sqrt{\frac{1-\lambda_0}{\lambda_0}}$. During the proof of Theorem 3.2.1, one sees that the first order terms do, in fact, behave like a kernel smoothing of iid $U(0,1)$ random variables. For these terms, the condition $(m \wedge n)h \to \infty$ is sufficient for asymptotic normality. The extra conditions, both on K and the rate at which h goes to 0, are necessary to show that the second order terms converge to zero in probability.

Although interesting in its own right, it is not clear how Theorem 3.2.1 might be used to test the null hypothesis. One certainly wouldn't base a test on the kernel estimate at a single point. It should be possible to extend Theorem 3.2.1 to show the joint convergence in distribution of $(\hat{d}_h(w_1), \ldots, \hat{d}_h(w_k))$ to a multivariate normal distribution for fixed k and $w_1, \ldots, w_k$. However, at this point one encounters a practical problem. To test H_o, how would one choose k and $w_1, \ldots, w_k$? There perhaps is some way to choose these points in an optimal fashion, however such a scheme would certainly depend on the true and unknown

values of the comparison density. Such statistics would also not fit the criteria outlined in Section 1 and Subsection 2.2.6.

Pursuing for the moment a test based on k values, a natural statistic to use would be of the form

$$\frac{1}{k}\sum_{i=1}^{k}[\hat{d}_h(w_i) - 1]^2.$$

Letting k grow large, a natural analogue to this statistic is

$$\int_0^1 [\hat{d}_h(w) - 1]^2 dw. \tag{3.2.1}$$

It is only appropriate that barring some reason to look at a specific and fixed set of w's that all of them should be considered. Statistics such as (3.2.1) cannot be handled by pointwise convergence in distribution results such as Theorem 3.2.1. Instead, weak convergence results are required. These are treated in Subsection 3.2.3. Although (3.2.1) is also a statistic and so not does not fit the criteria which have been outlined, a study of (3.2.1) leads to a methodology which does. Construction of tests is taken up in Subsection 3.3.

Another type of asymptotic result which is often of interest is consistency. The following theorem can be proved regarding the consistency of $\hat{d}_h(w)$.

Theorem 3.2.2. *Let $h \to 0$, $(m \wedge n)h^2 \to \infty$ as $m \wedge n \to \infty$ and let K have support $[-1,1]$ with K differentiable on $(-1,1)$. If $\lambda_{(N)} = \lambda_0$ is not a function of N or $d_{(N)}$ converges to d_o uniformly, then $\hat{d}_h(w) \xrightarrow{P} d_o(w)$ as $m \wedge n \to \infty$, otherwise $\hat{d}_h(w) - (1/h)\int_0^1 K[(w-u)/h]d_{(N)}(u)du \xrightarrow{P} 0$ as $m \wedge n \to \infty$.*

Consistency is generally regarded as a good property. Theorem 3.2.2 states that if h tends to zero at the appropriate rate then $\hat{d}_h(w)$ will indeed be consistent.

Again, it is interesting to compare Theorem 3.2.2 with results from ordinary kernel density estimation. In the iid case, pointwise consistency is achieved under the conditions $h \to 0$, $Mh \to \infty$ and uniform consistency if $h \to 0$, $Mh^2 \to \infty$. The extra conditions on $\lambda_{(N)}$ and uniform convergence result from the fact that one is approximating a function which itself is changing with N. Hence, one needs either that it isn't changing with N (*i.e.* $\lambda_{(N)} = \lambda_0$) or uniform convergence. In

a very real sense, the proofs of theorems in the iid case are much easier because a good deal more is known than simply the weak convergence of the empirical process. It is not surprising, then, that results obtained in the iid case should be stronger.

There is one last property of the kernel estimate which should be examined. This is the question of invariance. In this context, invariance refers to whether it makes a difference which of the two samples is called the first sample. Expanding the notation for this purpose, let $D_\lambda^F(w) = FQ^H(w)$ be the comparison distribution function when the population with distribution function F is called the first sample. Similarly, let $D_\lambda^G(w) = GQ^H(w)$ be the comparison distribution function when the population with distribution function G is called the first sample. Let $d_\lambda^F(w)$ and $d_\lambda^G(w)$ be the corresponding comparison density functions. Let λ be the weight given to distribution function F (*i.e.* the probability of choosing population F or the fraction of the total sample represented by population F). Parzen (1983) shows that d_λ^F and d_λ^G satisfy

$$\lambda d_\lambda^F(w) + (1-\lambda)d^G\lambda(w) = 1$$

for all $w \in [0,1]$. The boundary kernel estimates of these quantities satisfy

$$\begin{aligned}
&\lambda_{(N)}\hat{d}_h^F(w) + (1-\lambda_{(N)})\hat{d}_h^G(w) \\
&= \frac{m}{N}\frac{1}{mh}\sum_{i=1}^{m} K_s\left(\frac{w-(R_i/N)}{h}\right) + \frac{n}{N}\frac{1}{nh}\sum_{i=1}^{n} K_s\left(\frac{w-(S_i/N)}{h}\right) \\
&= \frac{1}{Nh}\sum_{i=1}^{N} K_s\left(\frac{w-(i/N)}{h}\right) \\
&\to 1 \text{ as } N \to \infty \text{ and } Nh \to \infty
\end{aligned}$$

for all $w \in [0,1]$, where $s = s(w,h)$ and S_i is the rank of Y_i in the pooled sample. The last sum is a rectangular sum approximation to the integral $\frac{1}{h}\int_0^1 K_s[(w-u)/h]du$ which is 1. The above convergence is shown as part of the proof of Theorem 3.2.1 in Appendix B. Asymptotically, the boundary kernel estimate obeys the invariance property of the population quantities.

3.2.3. *The Kernel Density Process.* In this subsection, the kernel density process is defined and a theorem concerning its weak convergence under a fixed

bandwidth is stated. The implications of a fixed bandwidth are discussed and the covariance kernel of the limiting process under H_o is found and investigated.

The kernel density process, $\mathrm{KDP}_{N,h}(w)$, is defined as

$$\mathrm{KDP}_{N,h}(w) = \frac{1}{h}\int_0^1 K_s\left(\frac{w-u}{h}\right) d\mathrm{CD}_N(u), \tag{3.2.2}$$

where K_s is a boundary kernel and CD_N is as previously defined. The process $\mathrm{KDP}_{N,h}(w)$ is simply a centered and scaled version of $\hat{d}_h(w)$, that is,

$$\mathrm{KDP}_{N,h}(w) = \sqrt{N}[\hat{d}_h(w) - d_h(w)],$$

where $d_h(w) = \frac{1}{h}\int_0^1 K_s[(w-u)/h]d_{(N)}(u)du$ is a smoothing of the comparison density.

To allow for some flexibility, the theorem concerning the weak convergence of $\mathrm{KDP}_{N,h}$ is stated for any boundary kernel which obeys the following regularity conditions.

Regularity Conditions. *Let $K_s(w)$ be a family of boundary kernels indexed by $s \in [0,1]$. It is required that the derivative of $K_s(w)$ with respect to w exist for each s and that $K_s'(w)$ be continuous on $\mathbb{R}$. Define*

$$\theta(\delta) = \sup_{|x-y|<\delta} \frac{1}{h^2}\int_0^1 \left| K_s'\left(\frac{x-u}{h}\right) - K_{s'}'\left(\frac{y-u}{h}\right)\right| du,$$

where $s = s(x,h)$ and $s' = s(y,h)$. It is required that $\theta(\delta) \to 0$ as $\delta \to 0$.

Lemma 3.2.1 states the conditions under which the Gasser-Müller boundary kernel satisfies the Regularity Conditions.

Lemma 3.2.1. *Let $K(t)$ be a differentiable kernel with support $[-1,1]$ satisfying (1) K is continuous on $\mathbb{R}$ and (2) K' is continuous on $\mathbb{R}$. Then the Gasser-Müller boundary kernel based on K satisfies the Regularity Conditions.*

From Lemma 3.2.1 it follows that the Gasser-Müller boundary kernel cannot be based on just any kernel. For instance, the popular Epanechnikov kernel, which is a quadratic function, cannot be used. However, the biweight kernel,

which is a quartic kernel, can be used. The biweight will be used throughout this work when a specific kernel is required. That the Epanechnikov kernel cannot be used is not of great concern. From a practical standpoint, there is not much to choose from within the class of kernels having support $[-1,1]$ which are probability density functions, symmetric about zero and continuous on $\mathbb{R}$. Very similar results can be obtained from these kernels by altering the bandwidths. The differences are in higher order smoothness properties which are difficult to detect visually. However, the smoother kernel is required here for the proofs to go through.

The limiting process of $\mathrm{KDP}_{N,h}$ is KDP_h which is defined as

$$\mathrm{KDP}_h(w) = \frac{1}{h^2}\int_0^1 K_s'\left(\frac{w-u}{h}\right) L(u)du,$$

where L is the limiting process of CD_N as defined in Subsection 2.3.3. Before proving that KDP_h is, in fact, the limiting process of $\mathrm{KDP}_{N,h}$, its existence must be shown. One must show that the defining integral equation has some meaning. This result is given by Lemma 3.2.2.

Lemma 3.2.2. *The sample paths of the process* KDP_h *exist and are continuous with probability 1.*

Now that the needed Regularity Conditions have been established and the limiting process exists with probability 1, the stage is set for the main result of this subsection. Theorem 3.2.3 gives the weak convergence of $\mathrm{KDP}_{N,h}$ to KDP_h.

Theorem 3.2.3. *If the boundary kernel satisfies the Regularity Conditions then for fixed* h *one has*

$$\mathrm{KDP}_{N,h} \Rightarrow \mathrm{KDP}_h$$

in $(C[0,1], \mathcal{C}_\rho, \rho)$ *as* $n \wedge m \to \infty$.

The triple $(C[0,1], \mathcal{C}_\rho, \rho)$ is a probability triple. The set of continuous functions on $[0,1]$ is $C[0,1]$; ρ is the sup-norm; and $\mathcal{C}_\rho$ is the σ-field generated by the open balls.

There are several aspects of Theorem 3.2.3 which merit attention. First is that the limiting process is basically a kernel smoothing of the process $L(u)$. The second is that the result refers to fixed bandwidths. This is in contrast to standard kernel density estimation results which require the bandwidth to shrink to zero. However, one should note that although some work has been done, it is also not typical in ordinary kernel density estimation to treat the estimator as a stochastic process and to investigate its weak convergence. Bickel and Rosenblatt (1973) have done the major work in this area. They assume the data is iid according to a density f and that f satisfies various conditions such as it is twice differentiable with a bounded second derivative. They then treat the kernel density estimate on a bounded interval as a process and obtain weak convergence results under a shrinking bandwidth.

There are several arguments that can be employed to justify fixing the bandwidth in Theorem 3.2.3. The rationale in any context for letting the bandwidth tend toward zero is to remove the bias of the estimator. In terms of testing, the comparison density is uniform under the null hypothesis. It is not hard to show that the estimator, $\hat{d}_h(w)$, is asymptotically unbiased under H_o for fixed bandwidths. Observe:

$$\begin{aligned} \mathrm{E}[\hat{d}_h(w)] &= \frac{1}{mh}\sum_{i=1}^{m}\mathrm{E}\left[K_s\left(\frac{w-(R_i/N)}{h}\right)\right] \\ &= \frac{1}{Nh}\sum_{i=1}^{N}K_s\left(\frac{w-(i/N)}{h}\right), \end{aligned}$$

since each R_i is marginally uniform over $1,\ldots,N$ under H_o. Hence,

$$\mathrm{E}[\hat{d}_h(w)] \to \frac{1}{h}\int_0^1 K_s\left(\frac{w-u}{h}\right)du = 1$$

as $N \to \infty$. Thus $\hat{d}_h$ is asymptotically unbiased for fixed h. From an implementation standpoint, knowing results like $(m \wedge n)h^3 \to \infty$ is not much help in choosing a bandwidth. Recall the discussion of Subsection 2.4.6. One is either left to judge the fit by graphical standards or by a criterion such as least squares cross-validation. Essentially, h controls the tradeoff between bias and variance. Letting $h \to 0$ may make things work out asymptotically, but is of little help in

fixed samples. Finally, one can argue that the asymptotic approximation derived for fixed h is superior to that derived for $h \to 0$. To this end, an examination of the exact differences in the limiting distributions should prove useful. Under H_o, each has the same limiting mean but the variances are not the same. The variance for $h \to 0$ is given in Theorem 3.2.1; a variance formula from Theorem 3.2.3 needs to be derived.

Let $C_h(v,w)$ be the covariance kernel of KDP_h. The covariance kernel is defined as $C_h(v,w) = \mathrm{E}[\mathrm{KDP}_h(v)\mathrm{KDP}_h(w)]$ which has support on the unit square. The covariance kernel of $\mathrm{KDP}_h(w)$ under H_o can be derived and the result is given as Lemma 3.2.3.

Lemma 3.2.3. *The covariance kernel of* $\sqrt{\lambda_0/(1-\lambda_0)}\mathrm{KDP}_h(w)$ *under* H_o *is*

$$C_h(v,w) = \frac{1}{h^2}\int_0^1 K_s\left(\frac{v-u}{h}\right) K_{s'}\left(\frac{w-u}{h}\right) du - 1, \tag{3.2.3}$$

where $s = s(v,h)$ *and* $s' = s(w,h)$.

Although the formula for $C_h(v,w)$ looks somewhat messy, it is possible to obtain a closed form expression for it. No numerical integration is necessary. This formula is derived in Appendix B.

From the definition of the covariance kernel, it is obvious that $C_h(v,w) = C_h(w,v)$. Recall the relation of the boundary kernel for the left and right endpoints from Subsection 2.4.3, $K^r_{s'}(t) = K^l_s(-t)$. Using this relation and a change of variable in (3.2.3) it can be shown that $C_h(v,w)$ also satisfies

$$C_h(v,w) = C_h(1-v, 1-w).$$

These symmetries are visible in Figures 8 through 10. Each of these figures presents four perspective plots of the covariance kernel under H_o. Figure 8 pictures $C_h(v,w)$ for $h = 0.5$; Figure 9 for $h = 0.3$; and Figure 10 for $h = 0.1$. The graphs are truncated at ± 3 so that details are not obscured by a large dynamic range. In each figure, the covariance changes from its base level of -1 only if the two points are within two bandwidths of one another. For a large bandwidth (Figure 8, $h = 0.5$), one observes a very smooth tunnel-like appearance. For a

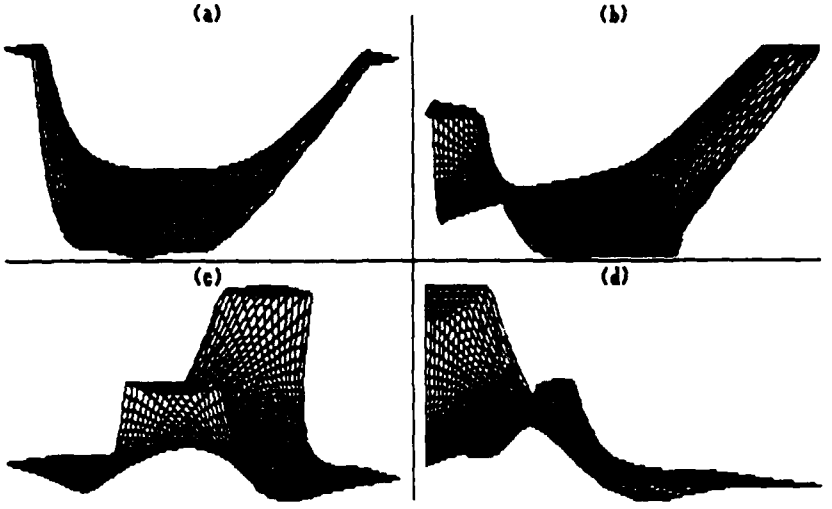

Fig. 8. *Perspective plots of the covariance kernel of the kernel density process under* H_o. *The bandwidth is* $h = 0.5$. *The perspective of Figure* (a) *is* $(5, -4, 10)$; *Figure* (b) *is* $(.5, -4, 10)$; *Figure* (c) *is* $(-2, -3, 10)$; *and Figure* (d) *is* $(0, -.5, 0)$.

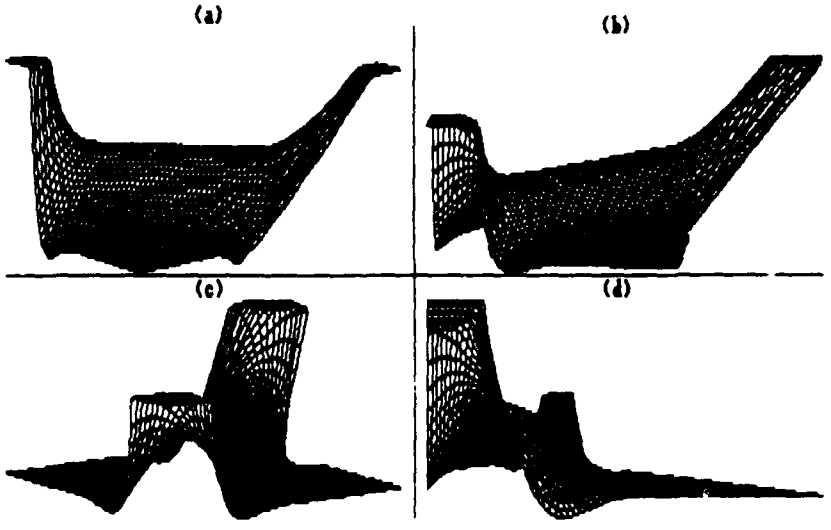

Fig. 9. *Perspective plots of the covariance kernel of the kernel density process under* H_o. *The bandwidth is* $h = 0.3$. *Th perspective of Figure* (a) *is* $(5, -4, 10)$; *Figure* (b) *is* $(.5, -4, 10)$; *Figure* (c) *is* $(-2, -3, 10)$; *and Figure* (d) *is* $(0, -.5, 0)$.

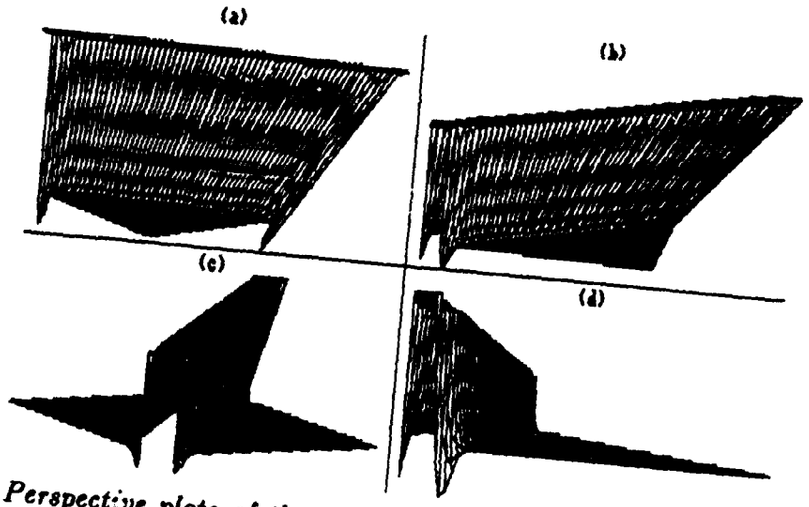

Fig. 10. *Perspective plots of the covariance kernel of the kernel density process under* H_o. *The bandwidth is* $h = 0.1$. *The perspective of Figure* (a) *is* $(5, -4, 10)$; *Figure* (b) *is* $(.5, -4, 10)$; *Figure* (c) *is* $(-2, -3, 10)$; *and Figure* (d) *is* $(0, -.5, 0)$.

small bandwidth (Figure 10, $h = 0.1$), one observes a very steep jump once two points are within two bandwidths of one another. The flat ridge results from truncating the plot though it's not hard to see that the variance increases with decreasing bandwidth.

The covariance kernel of the kernel density process can also be found under the alternative hypothesis. It is a rather complicated integral formula which depends not only on the boundary kernel but also on the comparison distribution function and the comparison density function. It is generally of little use, and so will not be given.

Returning to the question of fixed versus shrinking bandwidth for $\hat{d}_h(w)$, compare the differences in asymptotic variance formulas for the normalized random variable, $\sqrt{Nh}[\hat{d}_h(w) - 1]$ under H_o. For shrinking h, Theorem 3.2.1 gives the asymptotic variance as

(3.2.4)
$$\frac{1-\lambda_0}{\lambda_0}\int_0^1 K(t)^2 dt.$$

For fixed bandwidth, the variance formula is

(3.2.5)
$$\frac{1-\lambda_0}{\lambda_0}\Big[\int_{(w-1)/h}^{w/h} K_s(w)^2 dw - h\Big].$$

There are two distinctions between (3.2.4) and (3.2.5). For w near the boundary, (3.2.5) uses the boundary kernel whereas (3.2.4) does not. Equation (3.2.5) also has an additional term. Since (3.2.5) approaches (3.2.4) as h tends to zero, one could easily view (3.2.5) as containing correction factors for fixed h. Since h is fixed for finite samples, one might suppose that (3.2.5) would yield an improved approximation of the small sample distribution function by the limiting distribution function. A simulation confirms this supposition.

The simulation consists of 1000 replications for each of five choices of m and n. The two samples are each drawn from a $U(0,1)$ population, hence H_o is true. The bandwidth is given by $h = 0.75(m \wedge n)^{-.3}$ which satisfies the conditions of Theorem 3.2.1. For each replication, the comparison density is found for $w = 0.1, 0.2, 0.3, 0.4, 0.5$. Values above 0.5 are not needed by symmetry with those below 0.5. The one sample Kolmogorov-Smirnov statistic comparing the fit of the data to each normal approximation is then calculated for each sample. The Kolmogorov-Smirnov statistic is a measure of the goodness of fit of the sample and asymptotic distributions. The results are presented as Table 3. The Kolmogorov-Smirnov statistics based on the fixed bandwidth approximation are less than those for shrinking bandwidth in all but one case, $n = m = 100$ and $w = 0.1$. The difference in this case is at the second decimal place and is certainly statistically insignificant. The overall impression from Table 3 is that the Kolmogorov-Smirnov values for the fixed h approximation are substantially smaller than those for shrinking h. The lower values for the fixed h approximation would imply that this is the superior approximation.

The conclusion to draw from these remarks is that there are very good reasons to derive asymptotic results for fixed rather than shrinking bandwidths. The kernel density process forms the basis of tests of the null hypothesis. These tests are discussed in Subsection 3.3.

3.3. Tests of the Null Hypothesis

3.3.1. *Introduction.* In this subsection, tests of the null hypothesis, $H_o: d_\lambda(w) = 1$, are examined. The first test looked at is based on a statistic,

Table 3

Comparison of the fit of the small sample distribution of the Gasser-Müller boundary kernel estimate of the comparison density under H_o to its limiting distribution with fixed and shrinking bandwidth. The values in the table are one sample Kolmogorov-Smirnov statistics.

		w				
m	*n*	0.1	0.2	0.3	0.4	0.5
		Fixed Bandwidth				
20	20	3.45	1.60	0.74	0.99	0.78
50	20	3.70	1.57	0.71	1.43	1.23
50	50	1.52	0.81	1.27	0.61	1.17
50	100	1.30	0.70	0.56	1.05	1.42
100	100	1.73	0.67	1.00	0.79	0.60
		Shrinking Bandwidth				
20	20	4.31	2.19	2.03	1.90	1.97
50	20	4.39	2.52	2.39	2.05	2.26
50	50	1.65	2.08	1.93	1.81	1.88
50	100	1.39	1.57	1.52	1.98	1.96
100	100	1.70	1.82	1.94	1.14	1.72

$\hat{\varphi}^2_{N,h}$, which is a scaled version of the square of the $\mathcal{L}_2$ norm between $\hat{d}_h$ and 1. Since $\hat{\varphi}^2_{N,h}$ is a statistic, it does not fit the criteria required for a testing procedure. However, it leads to the concept of components similar to those discussed in Subsection 2.2.5. These components form the basis of the testing procedure. They are investigated in depth in Subsection 3.3. Their properties are explored and numerical methods for calculating the eigenfunctions and eigenvalues upon which they are based are examined. Subsection 3.3.4 looks at the question of whether the eigenfunctions form a complete basis for the spaces in which $KDP_{N,h}$ and KDP_h reside. This raises issues in terms of orthogonal decompositions of $\hat{d}_h$ by the eigenfunctions. Subsection 3.3.5 introduces a new test, the subset chi-square test, which is applied to the components in Subsection 3.3.6. Subsection 3.3.6 also introduces an orthogonal series estimator based on the components. The relation of the boundary kernel estimator and this orthogonal series estimator is investigated. Subsection 3.3.7 provides recommendations for the choice of

bandwidth. Finally, Subsection 3.3.8 summarizes a unified technique of estimation and testing which satisfies all the required criteria.

3.3.2. *The Statistic* $\hat{\varphi}^2_{N,h}$. In this subsection, the statistic $\hat{\varphi}^2_{N,h}$ is defined. Its limiting distribution is found and a possible representation for this limiting distribution in terms of components is also examined. Define $\hat{\varphi}^2_{N,h}$ by

$$\hat{\varphi}^2_{N,h} = \frac{\lambda_{(N)}}{1-\lambda_{(N)}} N \int_0^1 [\hat{d}_h(w) - 1]^2 dw,$$

which has already been briefly mentioned in Subsection 3.2.2. Note that 1 is subtracted from $\hat{d}_h$ and not $d_{(N)}$. The general statistic is not of interest because it is desired to test the null hypothesis and so the mean under H_o is subtracted. This will be the case throughout this subsection. Although the general weak convergence of $\mathrm{KDP}_{N,h}$ was shown in Subsection 2.2.3, for construction of tests one only needs the weak convergence under H_o. For clarity, the process $\mathrm{KDPo}_{N,h}$, which is defined as

$$\mathrm{KDPo}_{N,h} = \sqrt{N}[\hat{d}_h(w) - 1],$$

is introduced and will be referred to as the null kernel density process. The process $\mathrm{KDPo}_{N,h}$ equals $\mathrm{KDP}_{N,h}$ under H_o and so converges weakly under H_o. Under alternatives there is no such result: the process will become unbounded as N increases because it is incorrectly centered. This too is desirable as it is indicating that the null hypothesis is false. It is now possible to rewrite $\hat{\varphi}^2_{N,h}$ as $(\lambda_{(N)}/(1-\lambda_{(N)})) \int_0^1 \mathrm{KDPo}_{N,h}(w)^2 dw$.

The statistic $\hat{\varphi}^2_{N,h}$ is a normalized estimate of Pearson's φ^2 distance measure which can be written [cf. Eubank, LaRiccia, and Rosenstein (1987)] as

$$\varphi^2 = \int_0^1 [d_{(N)}(w) - 1]^2 dw.$$

The initial claim for $\hat{\varphi}^2_{N,h}$ is that it converges in distribution to the random variable

$$\varphi^2_h = \frac{\lambda_0}{1-\lambda_0} \int_0^1 \mathrm{KDP}_h(w)^2 dw$$

under H_o. This convergence in distribution can be demonstrated by results for functionals of stochastic processes which are known to converge weakly. It is not hard to show this functional is continuous [see Ruymgaart (1988), page 54]; it is also measurable $(C_\rho, \mathcal{B})$ where $\mathcal{B}$ is the Borel σ-field. Theorem 3.11 of Ruymgaart may be used to establish the fact that $\hat{\varphi}^2_{N,h} \xrightarrow{d} \varphi^2_h$. Having found the limiting random variable, φ^2_h, its distribution needs to be established.

The distribution of quantities such as φ^2_h has been examined in the literature. Its behavior is similar to that of the Cramér-von Mises statistic defined in Subsection 2.2.5. One would like to apply a technique similar to that applied by Durbin and Knott (1972) to the Cramér-von Mises statistic. That is, one would like to represent φ^2_h by

$$\varphi^2_h \sim \sum_{j=1}^{\infty} \theta^h_j Z^2_j, \tag{3.3.1}$$

where $Z_1, Z_2, \ldots$ are iid N(0,1) random variables and $\{\theta^h_j\}$ satisfies $\theta^h_j \geq 0$ and $\sum_{j=1}^{\infty} \theta^h_j < \infty$. The details of the basis of this representation can be found in Shorack and Wellner (1986). The Z_j's are known as components of the null kernel density process and the θ^h_j's will be seen to be eigenvalues of $C_h(v,w)$. The details of the construction of the components are not needed here but will be discussed in Subsection 3.3.3. The exact meaning of the '$\sim$' in equation (3.3.1) is subject to question. Under one set of conditions, it refers to 'distributed as'. However, under another set of conditions, the definitions of $\text{KDPo}_{N,h}$ and KDP_h need to be modified to be their projection onto an appropriate subspace. This projection is then substituted for the original process and the results hold. For the moment, these concepts and definitions are left intentionally vague. They will be discussed in depth in Subsection 3.3.4.

The statistic $\hat{\varphi}^2_{N,h}$ motivates the introduction of components as a natural consequence of investigating the distribution of the limiting random variable. It also motivates a detailed study of $C_h(v,w)$. From a practical standpoint the two interpretations do not matter much. However, which scenario holds will change certain interpretations and wordings. In the next subsection, the components will be seen to be of greater interest than the statistic which initially motivates

them.

3.3.3. *Components of the Kernel Density Process.* In this subsection the components of the kernel density process which were introduced above are defined. The basic properties of $C_h(v,w)$ are given first as they will be needed throughout. In order to define the components it is necessary to find the eigenfunctions and eigenvalues of $C_h(v,w)$. It is observed that the eigenfunctions and eigenvalues must be found numerically. A method for doing so is suggested. The resulting approximations are examined graphically. Finally, various interpretations of the components are explored. They are seen to be both generalized Fourier coefficients and linear rank statistics.

Before starting on a discussion of eigenfunctions, eigenvalues, and components a few of the properties of $C_h(v,w)$ are given. These will be needed throughout to establish the properties of these other objects of interest. These properties of $C_h(v,w)$ are stated as Lemma 3.3.1.

Lemma 3.3.1. *The covariance kernel $C_h(v,w)$ satisfies the following:*

i. *$C_h(v,w)$ is continuous on the unit square,*

ii. $\int_0^1 \int_0^1 C_h(v,w)^2 dvdw < \infty$,

iii. $\int_0^1 C_h(v,v)dv < \infty$.

The function $\phi^h(v)$, which is defined on $[0,1]$, is said to be an eigenfunction of $C_h(v,w)$ and θ^h is said to be the associated eigenvalue if $\phi^h(v) \not\equiv 0$ and

$$\int_0^1 \phi^h(v) C_h(v,w) dv = \theta^h \phi^h(w),$$

for all $w \in [0,1]$. Shorack and Wellner (1986), page 207, give a list of results for eigenfunctions and eigenvalues. Among these are:

1. The eigenvalues are at most countable in number.
2. Corresponding to any non-zero eigenvalue there are at most a finite number of linearly independent eigenfunctions; the maximal such number is called the multiplicity of the eigenvalue.
3. Let $\{\theta_j^h\}$ be an enumeration of the nonzero eigenvalues with each eigenvalue appearing as many times as its multiplicity. Then the set $\{\phi_j^h(v)\}$ of

eigenfunctions may be assumed to be orthonormal.

4. The eigenvalues $\{\theta_j^h\}$ satisfy $\int_0^1 C_h(v,v)dv = \sum_{j=1}^{\infty} \theta_j^h$.

The properties of $C_h(v,w)$ given in Lemma 3.3.1 imply several additional ones involving both $C_h(v,w)$ and the eigenfunctions. These are given as Lemma 3.3.2.

Lemma 3.3.2. *The eigenfunctions, $\phi_j^h(w)$, and the null covariance kernel, $C_h(v,w)$, satisfy:*

i. *The eigenfunctions, $\phi_j^h(w)$, are continuous on* $[0,1]$,

ii. $C_h(v,w)$ *is positive semi-definite,*

iii. $C_h(v,w) = \sum_{j=1}^{\infty} \theta_j^h \phi_j^h(v)\phi_j^h(w)$, *where the infinite series converges both absolutely and uniformly.*

The covariance kernel, $C_h(v,w)$, is said to be positive semi-definite if

$$\int_0^1 \int_0^1 g(v)C_h(v,w)g(w)dvdw \geq 0$$

for all $g \in \mathcal{L}_2[0,1]$ with $g \not\equiv 0$. It is said to be positive definite if the ≥ 0 can be replaced by > 0. These properties will all be of use at one point or another. For now the discussion turns to actually calculating the eigenfunctions and eigenvalues.

The form of $C_h(v,w)$ as given in equation (3.2.3) involves the boundary kernel in a very complicated way. It is probably too much to expect to be able to derive analytic expressions for the eigenfunctions and eigenvalues. This is indeed the case: a numerical solution is needed. The route taken is to approximate the defining integral by Simpson's rule and to convert the problem to an ordinary matrix eigenvalue problem. Such discretized approximations are well known in the literature; see, for example, Ahués, d'Almeida, Chatelin and Telias (1982).

It is desired to find $\phi_j^h(v)$ and θ_j^h to solve the integral equation

$$\int_0^1 \int_0^1 \phi_j^h(v)C_h(v,w)\phi_i^h(w)dvdw = \theta_j^h \delta_{ij}, \tag{3.3.2}$$

where δ_{ij} is Kronecker's delta. Equation (3.3.2) is approximated by a two dimensional Simpson's rule at the points $x^l = (0, 1/l, 2/l, \ldots, l/l)$ with l even. Define

$$d^l = (1, 4, 2, 4, \ldots, 2, 4, 1),$$

$$v'_{hj} = (\phi_j^h(0), \phi_j^h(1/l), \ldots, \phi_j^h(l/l)),$$

and

$$K_h = \Big[C_h\Big((i-1)/l, (j-1)/l\Big), \;\; i,j = 1,\ldots,l+1\Big].$$

The Simpson's rule approximation to equation (3.3.2) is

$$\frac{1}{9l^2} v'_{hj} D K_h D v_{hi} = \theta_i^h \delta_{ij}, \;\; i,j = 1,2,3,\ldots,$$

where $D = \text{diag}(d)$. Grouping these together for $i,j = 1,\ldots,l+1$ yields

$$\frac{1}{9l^2} V_h' D K_h D V_h = \Theta_h,$$

where $\Theta_h = \text{diag}(\theta_1^h,\ldots,\theta_{l+1}^h)$ and $V_h = [v_{h1},\ldots,v_{hl+1}]$. There is also an orthogonality condition on the ϕ_j^h's; namely

$$\int_0^1 \phi_j^h(w)\phi_i^h(w)dw = \delta_{ij}, \;\; i,j = 1,2,3,\ldots$$

The orthogonality condition is approximated by

$$\frac{1}{3l} v'_{hi} D v_{hj} = \delta_{ij}, \;\; i,j = 1,2,3,\ldots,$$

so in order to approximate the first $l+1$ eigenfunctions and eigenvalues, the following system must be solved for V_h and Θ_h:

$$\frac{1}{9l^2} V_h' D K_h D V_h = \Theta_h,$$
$$\frac{1}{3l} V_h' D V_h = I_{l+1}.$$

Letting $S_h = \frac{1}{\sqrt{3l}} D^{1/2} V_h$, this system is equivalent to

$$S_h' \Big[\frac{1}{3l} D^{1/2} K_h D^{1/2}\Big] S_h = \Theta_h,$$
$$S_h' S_h = I_{l+1}.$$

This last system is an ordinary symmetric eigenvalue problem which can be handled numerically. One solves it for S_h and Θ_h and then finds V_h from S_h.

Since the covariance kernel is so nicely behaved, one expects Simpson's rule to perform well in this case. One would also expect the eigenvalues to be better

estimated than the eigenfunctions; the former are simple scalars whereas the latter are functions. This, however, is a much more desirable state than the alternative. The eigenfunctions appear only in integrals where individual error is damped down, whereas it is often necessary to divide by the eigenvalues in which case error could produce large effects. In considering a choice for l, one should try to make it rather larger than the highest order eigenfunction one is considering using. Otherwise, the approximation may be at too few points to get a good fix on the function.

Applying this technique and using the biweight kernel and $l = 88$, Figures 11 through 13 are produced. Each figure presents, for a different h, the first four approximated eigenfunctions. Figure 11 is constructed with $h = 0.5$; Figure 12 with $h = 0.3$; and Figure 13 with $h = 0.1$. Generally, the j^{th} eigenfunction has a similar shape for each bandwidth up to an arbitrary sign. Changing the bandwidth tends to change the sharpness or peakedness of the functions. Notice that they are oscillatory; the j^{th} eigenfunction has j zero crossings in $(0,1)$. Figure 14 presents the estimated eigenvalues for these three bandwidths. The values all decay to zero as required considering they sum. The larger the bandwidth, the more quickly they decay to zero.

As a check on how well the numerical approximation works, one can compare the sum of the estimated eigenvalues with $\int_0^1 C_h(v,v)dv$. These two values should be comparable. The sum is truncated, but considering the rate at which the eigenvalues are decreasing this effect should be very small. Table 4 presents this comparison for five bandwidths, $h = 0.5, 0.4, 0.3, 0.2$, and 0.1. The true value is computed using Simpson's rule at 1001 points. The estimated values are astonishingly accurate. Interestingly, the estimated sum tends to err on the side of being slightly too big. The largest relative error is less than 0.03%, though it would be very slightly larger if one could add in the truncated values. Nevertheless, such good results are very encouraging. They lend credence to the approximating procedure as a whole.

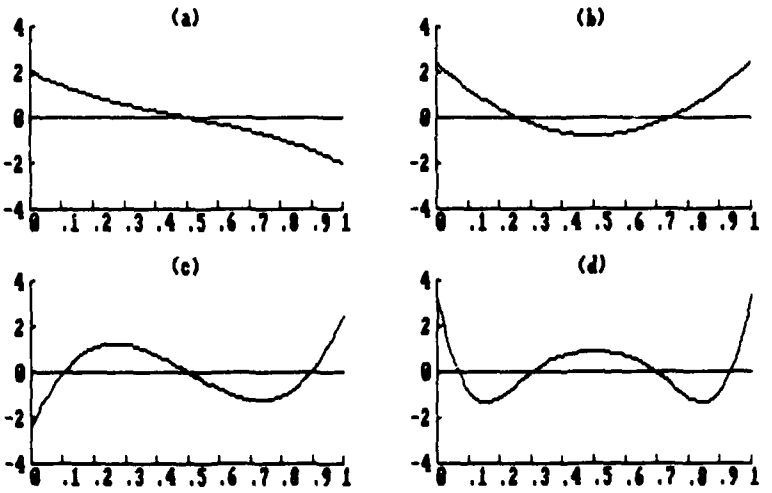

Fig. 11. *The first four approximated eigenfunctions for $h = 0.5$ and the Gasser-Müller boundary modification to the biweight kernel. Figure (a) is the first eigenfunction; Figure (b) the second; Figure (c) the third; and Figure (d) the fourth.*

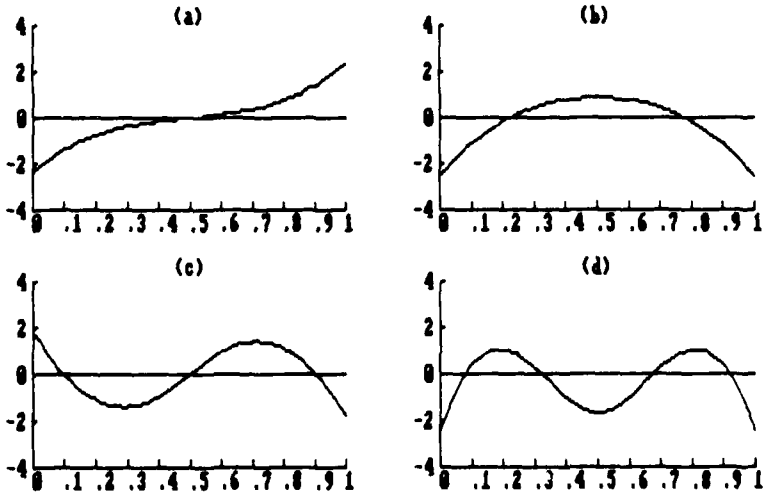

Fig. 12. *The first four approximated eigenfunctions for $h = 0.3$ and the Gasser-Müller boundary modification to the biweight kernel. Figure (a) is the first eigenfunction; Figure (b) the second; Figure (c) the third; and Figure (d) the fourth.*

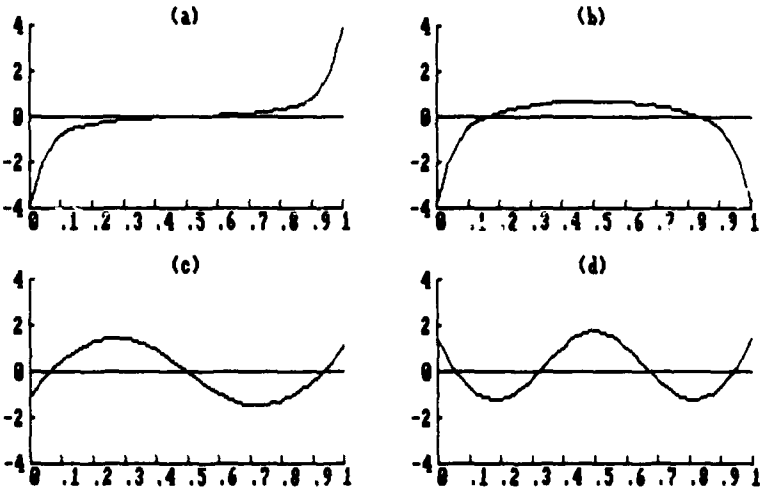

Fig. 13. *The first four approximated eigenfunctions for* $h = 0.1$ *and the Gasser-Müller boundary modification to the biweight kernel. Figure* (*a*) *is the first eigenfunction; Figure* (*b*) *the second; Figure* (*c*) *the third; and Figure* (*d*) *the fourth.*

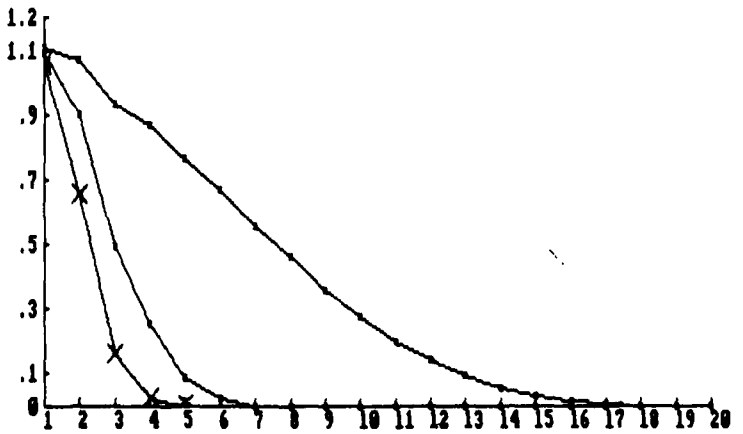

Fig. 14. *The first* 20 *estimated eigenvalues for* $h = 0.5, 0.3,$ *and* 0.1 *and the Gasser-Müller modification to the biweight kernel. The solid line with large* x*'s is* $h = 0.5$*; the solid line with blocks is* $h = 0.3$*; and the solid line with small* x*'s is* $h = 0.1$.

Table 4

Comparison of the sum of the estimated eigenvalues and their true sum.

	Sum of Eigenvalues	
h	True	Estimated
0.5	1.9074	1.9074
0.4	2.2646	2.2646
0.3	2.8598	2.8598
0.2	4.0503	4.0504
0.1	7.6217	7.6234

The components of the null kernel density process, $\mathrm{KDPo}_{N,h}$, are defined as

$$Z^*_{Nj} = \int_0^1 \phi_j^h(w)[\hat{d}_h(w) - 1]dw,$$

$$Z_{Nj} = \sqrt{\frac{\lambda_{(N)}N}{1-\lambda_{(N)}}} Z^*_{Nj}/\sqrt{\theta_j^h}$$

$$= \sqrt{\frac{\lambda_{(N)}}{1-\lambda_{(N)}}} \int_0^1 \phi_j^h(w)\mathrm{KDPo}_{N,h}(w)dw/\sqrt{\theta_j^h},$$

for $j \geq 1$ and the components of the limiting process are defined as

$$Z_j^* = \int_0^1 \phi_j^h(w)\mathrm{KDP}_h(w)dw$$

$$Z_j = \sqrt{\frac{\lambda_0}{1-\lambda_0}} Z_j^*/\sqrt{\theta_j^h},$$

for $j \geq 1$. Although the components clearly depend on h, it is not included in the notation for simplicity. By Lemma 3.3.1 and Proposition 2 of Shorack and Wellner (1986), page 208, one can conclude that $Z_1, Z_2, \ldots$ are iid N(0,1) random variables under H_o. Since the functional defining the components is continuous (as a result of Lemma 3.3.2) and is measurable $(C_\rho, \mathcal{B})$, by Theorem 3.11 of Ruymgaart (1988) and Theorem 3.2.3, one has

$$Z_{Nj} \xrightarrow{\mathrm{d}} Z_j$$

under H_o as $m \wedge n \to \infty$.

It has just been shown that the sample components converge in distribution singly to the appropriate limiting component. Later, the joint convergence in distribution of the set $Z_{N1}, \ldots, Z_{NM}$ will be needed. This result is easily derived by an application of the Cramér-Wold device and an appeal to the same theorems used above. Formally, Lemma 3.3.3 states the result.

Lemma 3.3.3. *For any fixed integer $M \geq 1$ and fixed bandwidth h, one has $(Z_{N1}, \ldots, Z_{NM}) \xrightarrow{\mathrm{d}} (Z_1, \ldots, Z_M)$ as $m \wedge n \to \infty$.*

The components have several very important interpretations. First, Z^*_{Nj} is the j^{th} generalized Fourier coefficient in an expansion of $\hat{d}_h(w) - 1$ in the eigenfunctions. This interpretation will be significant in a later subsection where an orthogonal series estimator of $d_{(N)}$ is based on the eigenfunctions. Second, Z^*_{Nj} is a linear rank statistic. This can be seen as follows:

$$\begin{aligned} Z^*_{Nj} &= \int_0^1 \phi_j^h(w)[\hat{d}_h(w) - 1]dw \\ &= \int_0^1 \phi_j^h(w) \left[\int_0^1 \frac{1}{h} K_s \left(\frac{w-u}{h} \right) dD_N(u) - 1 \right] dw \\ &= \int_0^1 \left[\int_0^1 \frac{1}{h} K_s \left(\frac{w-u}{h} \right) \phi_j^h(w) dw \right] dD_N(u) - \int_0^1 \phi_j^h(w) dw, \end{aligned}$$

where $s = s(w,h)$. This last quantity has the form of a linear rank statistic with the score function

$$J_j^h(u) = \frac{1}{h} \int_0^1 \phi_j^h(w) K_s \left(\frac{w-u}{h} \right) dw.$$

This score function can be termed a 'backward' smoothing of $\phi_j^h(w)$. The term backward is applied because the integration is with respect w and not u as is usual. The term $\int_0^1 \phi_j^h(w)dw$ is a centering constant and is equal to $\int_0^1 J_j^h(u)du$. For rank statistics, the centering constant arises naturally by centering $D_N(u)$ by the appropriate function, which under H_o is u. This converts $D_N(u)$ to $D_N(u) - u$ which is a multiplicative constant $(\sqrt{N})$ away from being the empirical comparison distribution process, CD_N, under H_o, that is,

$$\sqrt{N} Z^*_{Nj} = \int_0^1 J_j^h(u) d\mathrm{CD}_N(u)$$

$$= \sqrt{N} \int_0^1 J_j^h(u) d[D_N(u) - u]$$
$$= \sqrt{N} \left[\int_0^1 J_j^h(u) dD_N(u) - \int_0^1 J_j^h(u) du \right].$$

Figures 15 through 17 picture these score functions. In order to make them easier to view, they have been scaled so that each attains a maximum absolute value of 1. The important aspect of score functions is their shape, not their magnitude. Each figure presents four graphs corresponding to $j = 1, 2, 3, 4$. Figure 15 employs a bandwidth of $h = 0.5$; Figure 16 uses $h = 0.3$; and Figure 17 employs $h = 0.1$. The property of the number of zero crossings of the eigenfunctions is preserved by the score functions. Recalling the discussion of score functions in Subsection 2.2.4, the first component is seen to test location and the second scale. Higher order components are testing higher frequency departures of $d_{(N)}$ from uniformity. Although it would be nice to give these higher frequency departures moment interpretations such as skewness and kurtosis, such interpretations have not been demonstrated.

Score functions based on eigenfunctions and boundary kernels are entirely novel: this is a new procedure for generating score functions. There are several attractive features to this methodology. First, one is generating an entire family of score functions—starting with location, moving to scale, and then higher order departures. There is a link between these since they have a unified origin. Portmanteau tests for departures up to the fourth order score function, (*i.e.* $j = 4$), have been proposed in the literature [see Boos (1986)]. However, the score functions employed have no common origin making the entire procedure seem somewhat *ad hoc*. Second, these score functions are parametrically defined by the bandwidth, h. Selecting the bandwidth allows one to select the properties of the test, that is, one can tune the bandwidth so that the components are more powerful against certain classes of underlying distributions. Eubank, LaRiccia, and Rosenstein (1987) provide a unified origin for score functions by taking them to be an orthonormal basis. There are several reasons why an approach based on eigenfunctions is desirable. First, to protect against different classes of distributions requires a complete change of the basis. Such a change of basis may

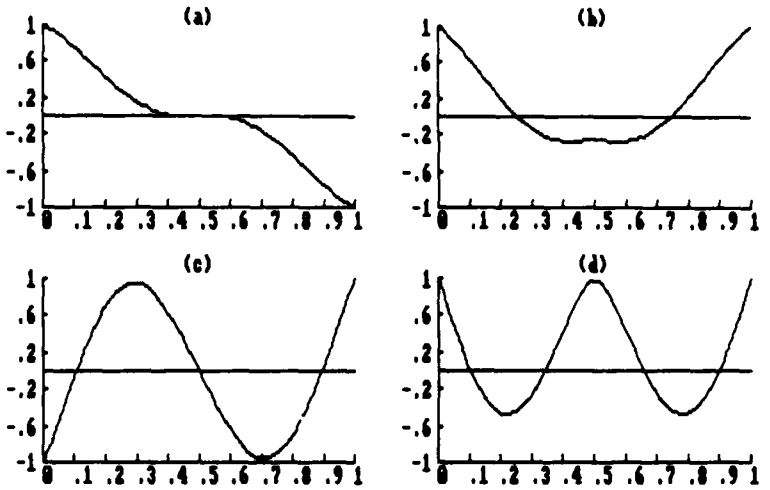

Fig. 15. *The first four score functions corresponding to the first four components of the kernel density process based on the Gasser-Müller boundary modification to the biweight kernel. The bandwidth is* $h = 0.5$. *The score functions have been scaled so that the maximum absolute value attained is* 1. *Figure* (*a*) *is the first score function; Figure* (*b*) *the second; Figure* (*c*) *the third; and Figure* (*d*) *the fourth.*

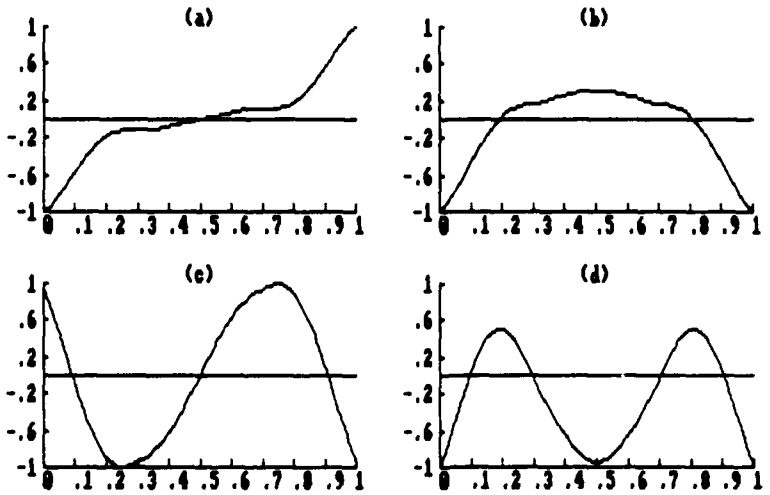

Fig. 16. *The first four score functions corresponding to the first four components of the kernel density process based on the Gasser-Müller boundary modification to the biweight kernel. The bandwidth is* $h = 0.3$. *The score functions have been scaled so that the maximum absolute value attained is* 1. *Figure* (*a*) *is the first score function; Figure* (*b*) *the second; Figure* (*c*) *the third; and Figure* (*d*) *the fourth.*

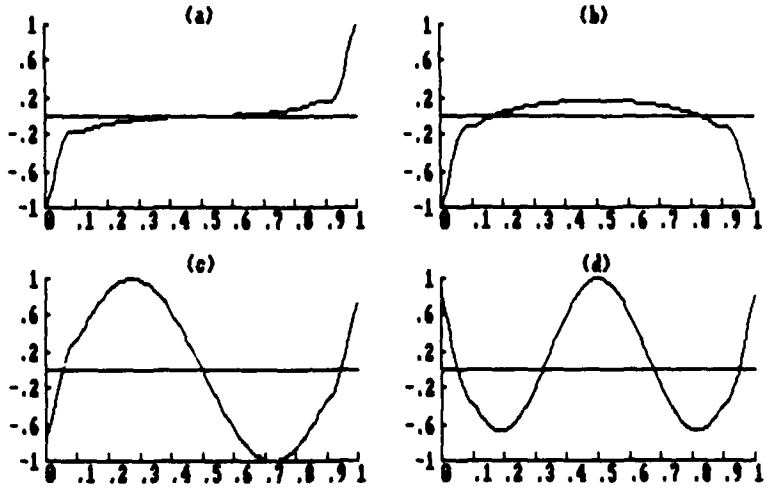

Fig. 17. *The first four score functions corresponding to the first four components of the kernel density process based on the Gasser-Müller boundary modification to the biweight kernel. The bandwidth is* $h = 0.1$. *The score functions have been scaled so that the maximum absolute value attained is* 1. *Figure* (*a*) *is the first score function; Figure* (*b*) *the second; Figure* (*c*) *the third; and Figure* (*d*) *the fourth.*

have important implications for any estimator of the comparison density based on it (cf. Subsection 2.4.4). Second, the eigenfunction-based scores have quite unusual shapes that would be hard to match by standard orthogonal functions; for instance, $J_1^{0.1}(u)$ puts more weight on the tails than would be observed for the Legendre polynomials or trigonometric functions. Often, the tails are precisely the area of interest. Thus, the eigenfunction approach generates interesting shapes that would be difficult to obtain otherwise. There is great convenience and theoretical unity in a procedure based on the eigenfunctions.

The asymptotic relative efficiencies (ARE's) of the components relative to standard two sample tests are taken up in detail in Section 4. There is a certain amount of ground work that needs to be performed in order to define the ARE's for the components. This work is most properly done in Section 4. Suffice it to say for now that altering the bandwidth does truly affect the properties of the components. These score functions give one the ability to choose the tests in a unified manner.

Rank statistics enjoy a very important invariance property. When properly centered and scaled, it doesn't matter which sample is called the first sample. The resulting rank statistics will have the same magnitude (but different signs). Hence, tests based on either statistic will always reach the same conclusion. Unfortunately, the rank statistic Z_{Nj} is not properly centered in small samples for this to be the case. As shown above, Z_{Nj} is centered by its asymptotic mean, $\int_0^1 J_j^h(u)du$. Asymptotically, everything works out fine. In finite samples, $\int_0^1 J_j^h(u)du$ can be sufficiently different from the small sample mean that invariance is lost. In fact, if m and n are very different, invariance may be lost to the point that one may reach different conclusions a significant amount of the time.

To demonstrate these statements, let $U^* = \frac{1}{m}\sum_{i=1}^m J_j^h(R_i/N)$ be the rank statistic where R_i is the rank of X_i in the pooled sample. Under H_o,

$$\begin{aligned} \mathrm{E} &\equiv \mathrm{E}[U^*] \\ &= \mathrm{E}\left[\frac{1}{m}\sum_{i=1}^m J_j^h(R_i/N)\right] \\ &= \frac{1}{N}\sum_{i=1}^N J_j^h(i/N), \end{aligned}$$

since each R_i is marginally distributed as uniform over $1,\ldots,N$. Let

$$\begin{aligned} U &= \sqrt{\frac{N\lambda_{(N)}}{1-\lambda_{(N)}}}(U^* - E) \\ &= \sqrt{\frac{mN}{n}}(U^* - E). \end{aligned}$$

Now, it is true that $\sqrt{N}(E - \int_0^1 J_j^h(u)du) \to 0$ as $N \to \infty$. This is shown as part of the proof of Theorem 3.2.1 in Appendix B. By Lemma A on page 20 of Serfling (1980), U has the same limiting normal distribution as Z_{Nj}. Let $V^* = \frac{1}{n}\sum_{i=1}^n J_j^h(S_i/N)$ where S_i is the rank of Y_i in the pooled sample. It follows that $\mathrm{E}[V^*] = E$. Let

$$\begin{aligned} V &= \sqrt{\frac{(1-\lambda_{(N)})N}{\lambda_{(N)}}}(V^* - E) \\ &= \sqrt{\frac{nN}{m}}(V^* - E). \end{aligned}$$

The random variable V has the same limiting normal distribution as U under H_o. Simple algebra yields the invariance result: $U + V = 0$. Instead of centering by E, suppose $F = \int_0^1 J_j^h(u)du$ is used. In this case, the random variables U and V no longer have mean zero but are biased. As noted above, this bias disappears asymptotically but may be significant in finite samples. The means of U and V are

$$\mathrm{E}[U] = \sqrt{\frac{mN}{n}}(E - F),$$
$$\mathrm{E}[V] = \sqrt{\frac{nN}{m}}(E - F).$$

If, say, $m \gg n$, then U will be considerably more biased than V. Comparing each to a standard normal reference distribution, one would expect U to reject more often than V under H_o. To avoid this problem and to preserve invariance, the small sample means will be subtracted throughout. The asymptotics are unaffected and the small sample properties improved.

The small sample mean of U^* is

$$\frac{1}{N}\sum_{i=1}^{N} J_j^h(i/N) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}\int_0^1 \phi_j^h(w)K_s\left(\frac{w - i/N}{h}\right)dw$$
$$= \int_0^1 \phi_j^h(w)\left[\frac{1}{Nh}\sum_{i=1}^{N} K_s\left(\frac{w - i/N}{h}\right)\right]dw.$$

This last formula is a more tractable form for calculation. The quantity inside the brackets is the kernel smoothing of the data $(1/N, 2/N, \ldots, N/N)$. The integral can be evaluated by Simpson's rule at $l + 1$ points (l even), hence one needs the kernel smoothing of $(1/N, 2/N, \ldots, N/N)$ at these $l + 1$ points. This can be calculated using the same routine that calculates the estimate of the comparison density.

As noted in the introduction, it is planned to use the components as the basis of the integrated testing and estimation procedure. The exact procedure has yet to be introduced; however a justification for using components can be made at this point. Even if statistics such as $\hat{\varphi}^2_{N,h}$ fit the criteria outlined, the components would still be of greater interest. A test based on components examines the first

M of them and gives each equal weight. This is in contrast to statistics such $\hat{\varphi}^2_{N,h}$ or the Cramér-von Mises which use all the components but downweight each successively according to the eigenvalues of the covariance kernel. This is clear from representations such as (3.3.1) and (2.2.3). Because they employ all the components, such statistics are consistent against any alternative. By this is meant that the probability of rejecting H_o if H_o is false tends to 1 as $m \wedge n \to \infty$. There is a price to be paid for this consistency, however. The weights on the components drop off so quickly that it takes a tremendous amount of data to detect an alternative which effects one of the higher order components [cf. Randles and Wolfe (1979), page 383]. This will be seen to be the case in Section 4. Such statistics begin to lose power for alternatives affecting even the second or third component.

In contrast, a procedure which tests only the first M components and gives each equal weight should have good power characteristics against alternatives affecting these components. However, it will be inconsistent against alternatives which effect only components other than the M considered. Such a tradeoff seems a reasonable one on several grounds. First, since the statistics will be seen to have poor power against even low order components, consistency is not much solace. Second, since M is under the control of the user, it can be chosen to protect against as broad a class as is felt necessary or suitable. One also has the comfort that this class is much better protected against than by the standard statistics.

In summary, this subsection has defined the basic machinery necessary to define the components. Properties of the covariance kernel of the null kernel density process have been investigated. The eigenfunctions and eigenvalues of this covariance kernel have been defined and their properties explored. A numerical method of finding these was suggested. The components were defined and given several very important interpretations. The components are both generalized Fourier coefficients and linear rank statistics. As linear rank statistics they are seen to be testing successively higher frequency departures of $d_{(N)}$ from uniformity. A small sample correction to the mean of the components was suggested

to preserve invariance. Finally, a first comparison of a test based on components versus the usual type of portmanteau statistic was made. It was argued that one would expect the components to have superior power against many alternatives. This will be seen conclusively in Section 4.

3.3.4. *The Space Spanned by the Eigenfunctions.* In this subsection, the space spanned by the eigenfunctions of the null covariance kernel is examined. The properties of decompositions based on them are also explored. It is seen that from a practical aspect, this question is of little import. However, from a theoretical aspect certain interpretations and representations do change. The question of exactly what space the eigenfunctions do span is not resolved. Given that there is no explicit representation for the eigenfunctions, their properties are all the more difficult to determine.

In deriving representations such as (3.3.1), it is necessary to know whether the eigenfunctions form a complete orthonormal basis for the space in which the stochastic process resides. Since it will be necessary to decompose both $\mathrm{KDPo}_{N,h}$ and KDP_h, an examination of the spaces in which these processes reside is necessary. Lemma 3.2.2 states the KDP_h is continuous with probability 1. A more precise statement than this is possible. The process KDP_h is in the space S_C^h with probability 1, where

$$S_C^h = \{f : f(w) = \int_0^1 \frac{1}{h^2} K_s'\left(\frac{w-u}{h}\right) g(u)du, s = s(w,h), g \in C[0,1]\}.$$

The process $\mathrm{KDPo}_{N,h}$ is in the space S_D^h with probability 1, where

$$S_D^h = \{f : f(w) = \int_0^1 \frac{1}{h^2} K_s'\left(\frac{w-u}{h}\right) g(u)du, s = s(w,h), g \in D[0,1]\},$$

and $D[0,1]$ is the space of all functions on $[0,1]$ which are continuous from the right and have limits from the left. From the definitions, it follows that $S_C^h \subset S_D^h$ so that the fundamental question is whether the eigenfunctions form a complete orthonormal basis for S_D^h. If they do, then the following two results follow:

$$\|\hat{d}_h - 1 - \sum_{j=1}^{M} Z_{Nj}^* \phi_j^h(w)\| \xrightarrow{\text{a.s.}} 0 \text{ as } M \to \infty, \tag{3.3.3}$$

$$\varphi_h^2 \stackrel{\mathrm{d}}{=} \sum_{j=1}^{\infty} \theta_j^h Z_j^2, \tag{3.3.4}$$

where $\| \cdot \|$ is the $\mathcal{L}_2$ norm, $\stackrel{\mathrm{d}}{=}$ denotes equal in distribution and $\xrightarrow{\text{a.s.}}$ denotes almost sure (with probability 1) convergence. See Shorack and Wellner (1986), pages 270 ff., for demonstrations of these facts. Basically, (3.3.3) is a statement of completeness and (3.3.4) is a form of Parseval's theorem.

If the eigenfunctions are not complete for S_D^h then equation (3.3.3) and (3.3.4) must be amended slightly. Let S_S^h be the space actually spanned by $\{\phi_j^h\}$. Let $\mathrm{P}\hat{d}$ be the projection of $\hat{d}_h(w) - 1$ onto S_S^h and PKDP_h be the projection of KDP_h onto S_S^h. Instead of (3.3.3) and (3.3.4), one has

$$\|\mathrm{P}\hat{d}(w) - \sum_{j=1}^{M} Z_{Nj}^* \phi_j^h(w)\| \xrightarrow{\text{a.s.}} 0 \text{ as } M \to \infty \tag{3.3.5}$$

$$\frac{\lambda_0}{1-\lambda_0} \int_0^1 \mathrm{PKDP}_h(w)^2 dw \stackrel{\mathrm{d}}{=} \sum_{j=1}^{\infty} \theta_j^h Z_j^2. \tag{3.3.6}$$

The results are the same, but now one must deal with the projections instead of the original processes.

This is inconvenient mathematically; practically it makes no difference. It makes no difference because when (3.3.3) and (3.3.4) are actually applied, they are truncated at some point, M. Hence, one is always dealing with the projection onto a subspace; for instance, in Subsection 3.3.6 the orthogonal series estimate of $d_{(N)}$,

$$\hat{d}_{h,M}(w) = 1 + \sum_{j=1}^{M} Z_{Nj}^* \phi_j^h(w),$$

is introduced as a truncated decomposition of $\hat{d}_h(w) - 1$. Since M is always finite, what does or does not happen in the tail of the sequence is of little importance. This is particularly true since the components, Z_{Nj}^*, are becoming small with high probability for increasing j. Note that Z_{Nj}^* is $\mathrm{AN}(0,(1-\lambda_0)\theta_j^h/(N\lambda_0))$ under H_o. Because $\theta_j^h \to 0$ as $j \to \infty$, the Z_{Nj}^*'s are becoming small with high

probability as j increases. The higher order eigenfunctions simply do not carry much weight.

In Section 4, the approximate distribution of φ_h^2 under H_o is found by numerically inverting the characteristic function of

$$\sum_{j=1}^{M} \theta_j^h Z_j^2. \tag{3.3.7}$$

One chooses M sufficiently large so that θ_j^h is very small for $j > M$. The percentage points of (3.3.7) are then used in proxy for those of φ_h^2. This is the methodology employed by Durbin and Knott (1972) with excellent results. The calculated distribution does not depend on whether the tail of the sequence fills out to be complete for S_D^h or not.

Shorack and Wellner (1986) state that a necessary and sufficient condition for the eigenfunctions to span a given space and for the eigenvalues to all be positive is that $C_h(v,w)$ be positive definite over that space. Here, it is required that

$$\int_0^1 \int_0^1 f(v) C_h(v,w) f(w) dv dw > 0,$$

for all $f \in S_D^h$ with $f \not\equiv 0$. Lemma 3.3.2 states that $C_h(v,w)$ is positive semi-definite over this space. Checking positive definiteness turns out to be no small task. Lemma 3.3.4 gives an equivalent condition.

Lemma 3.3.4. *$C_h(v,w)$ is positive definite on S_D^h if and only if the integral equation,*

$$\frac{1}{h} \int_0^1 K_s\left(\frac{w-u}{h}\right) f(w) dw = c,$$

has no solution for $f \in S_D^h$ where $f \not\equiv 0$ and $c = 0, 1$.

Unfortunately, the condition of Lemma 3.3.4 is no easier to check than the initial statement of positive definiteness. The integral equation of Lemma 3.3.4 is a Fredholm integral equation of the first kind. These are among the hardest to solve both analytically and numerically; see, for example, Marti (1982) and

Lukas (1980). Further, the integrand involves the boundary kernel and so is a rational function of w; it is not even a 'nice' Fredholm equation of the first kind. There seems no hope of an analytic solution. Numerical approaches are doubly difficult in this context since the question is over the existence of a solution, not finding a solution which is known to exist. Hence, one would be put in the position of trying to gauge whether the approximation is converging to a true solution or not. A further difficulty of numerical approaches is restricting the solution to reside in the space S_D^h. Such a restriction would be very difficult to impose.

The question of exactly what space the eigenfunctions span will not be resolved here. Where appropriate, the implications of the eigenfunctions forming a complete basis for S_D^h or failing to do so will be noted.

3.3.5. *The Subset Chi-Square Test.* In this subsection, a new test is presented. This test is referred to as the subset chi-square test. It is applied to the components, $Z_{N1}, \ldots, Z_{NM}$, in Subsection 3.3.6. It is seen to have several desirable properties. First, it represents a compromise between two existing tests: the standard chi-square test and the independent tests method. This latter tests the components one at a time at a smaller size than desired so that the overall test has the desired size. Second, the subset chi-square test indicates not only that some components are significant when it rejects, but also which ones are significant. Third, it lends itself to graphical display of a criterion function much in the way Akaike's (1974) AIC and Parzen's (1977) CAT criteria do in time series analysis.

Let $S_1, \ldots, S_M$ be independent normally distributed random variables with variance 1 and suppose that S_i has mean μ_i. A test of $H_o: \mu_1 = \mu_2 = \cdots = \mu_M = 0$ versus $H_a: \mu_i \neq 0$ for at least one i is desired. The need to conduct tests of this nature is not unknown in statistics. For instance, in time series analysis under the null hypothesis of no autocorrelation, the first M standardized sample autocorrelations are asymptotically iid N(0,1) [see Newton (1988), page 158]. The two most commonly used tests in this framework are the chi-square test and the independent tests method.

There do not exist optimal tests such as the uniformly most powerful unbiased test in this case–the alternatives are just too general. One can, of course, form several classically motivated statistics: likelihood ratio, Wald, Rao's efficient score can all be derived. In this situation, one usually sees one of two tests applied. The first is the chi-square test statistic

$$T = \sum_{i=1}^{M} S_i^2,$$

which is distributed as χ^2_M under H_o. The second is to test each of the S_i's one at a time and adjust the size of each test so that the overall test has the desired size, α. This is termed the independent tests method. In this procedure, H_o is rejected if and only if

$$S_i^2 > \chi_1^2((1-\alpha)^{1/M}),$$

for any $i = 1, \ldots, M$, where $\chi_1^2((1-\alpha)^{1/M})$ is the quantile of the χ_1^2 distribution evaluated at $(1-\alpha)^{1/M}$. The overall test does indeed have size α.

There are several considerations in choosing a test. The first is that it should have reasonable power against a wide range of alternatives. A second consideration is that if the null hypothesis is rejected, the test should indicate why it has been rejected. That is, it should indicate which of the S_i's were judged to have non-zero means. For the components, this information is useful on two grounds. First, knowing which rank statistic is large serves as a numerical indicator to accompany the estimate of the comparison density. Graphs should always be accompanied by diagnostics to reinforce the message. Knowing that the second component is the major difference between the samples is meaningful. Second, it will be seen that one can construct an orthogonal series estimate of $d_{(N)}$ based on the significant components. This follows from their interpretation as generalized Fourier coefficients. Finally, it would also be desirable if the test has some graphical components in the manner of AIC or CAT. One would like something more than just a list of significant components.

What is needed is a unified framework in which to discuss these two tests and to derive new ones. An appropriate framework is suggested by analogy with

the optimal subset regression techniques of Furnival (1971) and Furnival and Wilson (1974). The idea behind optimal subset regression is to find that subset of a given size, k, of the regressors which yields the least RSS (residual sum of squares). One can then vary k and use some criterion function such as Mallow's C_p to help choose a subset size. The key concept here is the examination of all subsets to yield the best result. By analogy, one could look at all subsets of size k of $S_1^2, \ldots, S_M^2$ for $k = 1, \ldots, M$ and reject H_o if the sum of the members of some subset were found to be too large. The philosophy is that $S_1^2, \ldots, S_M^2$ should not be considered only singly but should also be allowed to reinforce one another.

Mathematically, such a test works out to be: Reject H_o if and only if

$$S_{i_1}^2 + \cdots + S_{i_k}^2 > D^M(k, \alpha),$$

for some k, $1 \leq k \leq M$ and some $(i_1, \ldots, i_k)$. The indices $(i_1, \ldots, i_k)$ range over all $\binom{M}{k}$ subsets of size k taken from $\{1, \ldots, M\}$. The sequence $D^M(1, \alpha), \ldots, D^M(M, \alpha)$ are critical values which must be selected. This sequence determines the properties of the test and must keep the overall size at α.

To perform this test, one needn't actually look at all $2^M - 1$ subsets. In fact, even the branch and bound algorithm of Furnival and Wilson is unnecessary. All one need examine are M subsets. The above test is equivalent to: Reject H_o if and only if

$$S_{(M)}^2 + \cdots + S_{(M-k+1)}^2 > D^M(k, \alpha),$$

for some k, $1 \leq k \leq M$, where $S_{(1)}^2, \ldots, S_{(M)}^2$ are the order statistics of $S_1^2, \ldots, S_M^2$. There is no onerous computational burden at all. However, the optimal subset analogy is more motivational than simply starting with the order statistics.

At first blush, there seem to be at least three reasonable choices for the sequence of critical values, $D^M(1, \alpha), \ldots, D^M(M, \alpha)$. These are:

1. $D^M(k, \alpha) = \chi_M^2(1 - \alpha)$;
2. $D^M(k, \alpha) = k\chi_1^2((1 - \alpha)^{1/M})$;

3. $D^M(k,\alpha) = \chi^2_k(t(M,\alpha))$.

These shall be referred to as sequences (1), (2), and (3), respectively. Sequence (1) yields the ordinary chi-square test. For the chi-square test, the null hypothesis is rejected if $\sum_{i=1}^{M} S_i^2 > D^M(M,\alpha) = \chi^2_M(1-\alpha)$. If the chi-square rejects, the test based on sequence (1) will, also. Suppose, for some subset of size k, $(i_1,\ldots,i_k)$, one has

$$S_{i_1}^2 + \cdots + S_{i_k}^2 > D^M(k,\alpha).$$

Then, of course,

$$\sum_{i=1}^{M} S_i^2 \geq S_{i_1}^2 + \cdots + S_{i_k}^2 > D^M(k,\alpha) = \chi^2_M(1-\alpha),$$

so the chi-square test rejects as well. Thus, sequence (1) is equivalent to the standard chi-square test.

Sequence (2) is equivalent to the independent tests method. Clearly, if the independent tests method rejects then the test based on (2) will, also. Suppose, for some k and some subset $(i_1,\ldots,i_k)$ that

$$S_{i_1}^2 + \cdots + S_{i_k}^2 > k\chi^2_1((1-\alpha)^{1/M}).$$

If $S_{i_j}^2 < \chi^2_1((1-\alpha)^{1/M})$ for $j = 1,\ldots,k$ then the above cannot hold. Hence $S_{i_j}^2 > \chi^2_1((1-\alpha)^{1/M})$ for some j and the independent tests method also rejects. Thus sequence (2) is equivalent to the independent tests method.

The critical sequence (3) yields what is to be called the subset chi-square test. The critical value is a natural one since, taken alone, each term $S_{i_1}^2 + \cdots + S_{i_k}^2$ is distributed as χ^2_k under H_o. To keep the overall test at size α, one needs to adjust the size of each term in the critical sequence. This is the purpose of evaluating the χ^2_k quantile at the point $t(M,\alpha)$. All three tests are now in a common framework so that comparisons are possible.

Figure 18 represents the critical regions for the chi-square test and the independent tests method for $M = 2$. The square is the critical region for the independent tests method and the circle is the critical region for the chi-square

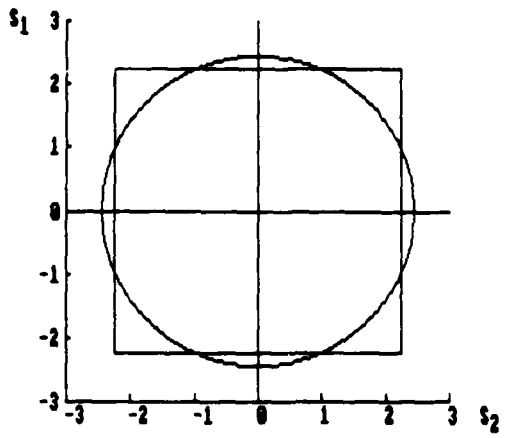

Fig. 18. *Critical regions for the chi-square test and the independent tests method for* $M = 2$ *and* $\alpha = 0.05$. *The critical region for the chi-square test is the circle and that for the independent tests method is the square.*

test. These regions are for a size of 0.05. The tests reject when the pair (S_1, S_2) falls outside the appropriate figure. Comparing the two, it is possible to draw some tentative conclusions about which test is more powerful for certain kinds of alternatives. If the alternative is in the direction of one of the axes, the independent tests method should outperform the chi-square test because its critical region is shorter in that direction. Similarly, if the alternative is in the direction of one of the diagonals, the chi-square test should do better.

Figure 19 presents the critical region for the subset chi-square test. The test rejects anytime the pair (S_1, S_2) falls outside either the circle or the square. Comparing the dimensions of these shapes to those in Figure 18, one sees those in Figure 19 are slightly larger. Visually, the subset chi-square test appears to be a compromise between the other two. It is a compromise of the independent tests method by cutting off the corners of the square. It is a compromise of the chi-square test by reducing the distance in the direction of the axes. If the alternative is in the direction of an axis, the independent tests method and the subset chi-square should be about the same and both better than the standard chi-square test. If the alternative is in the direction of a diagonal, the chi-square test should

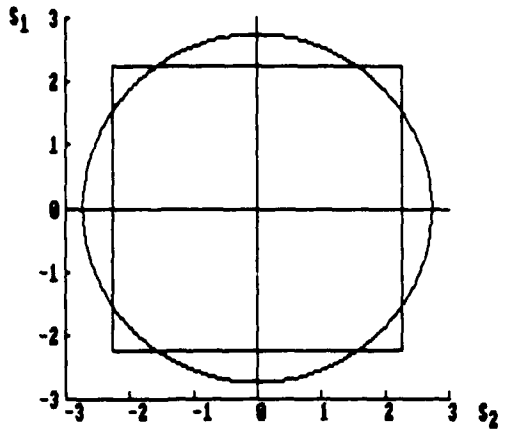

Fig. 19. *Critical region for the subset chi-square test for* $M = 2$ *and* $\alpha = 0.05$. *The test rejects any time the observation* (S_1, S_2) *falls outside either the square or the circle.*

perform the best followed by the subset chi-square and the independent tests method.

A small power study confirms these findings. Figure 20 presents the power curves of the three tests for $M = 4$ and alternative $(\mu_1, \mu_2, \mu_3, \mu_4) = q \cdot (1, 1, 1, 1)$. The scalar q ranges from 0 to 6.4. The power curves are constructed by simulation techniques using 10,000 replications. For each replication, four independent normal random variables are drawn with mean 0 and variance 1. For this realization, a loop steps through the range of q values. For each value q_j of q the appropriate means are added to the normal random variables drawn above. The sample power curve for the i^{th} test at the alternative q_j is calculated as 1 if the test rejects and 0 otherwise. The estimates are then averaged over all realizations to generate Figure 20. Clearly, there is reuse of each sample over the range of alternatives. However, this technique will converge to the correct values and it does impose monotonicity on the estimated power functions. From Figure 20, one can see that the ordering of the tests is as predicted.

Figure 21 repeats the same method for the alternative $(\mu_1, \mu_2, \mu_3, \mu_4) = q \cdot (1, 0, 0, 0)$. Again the results are as predicted, however the curves are much

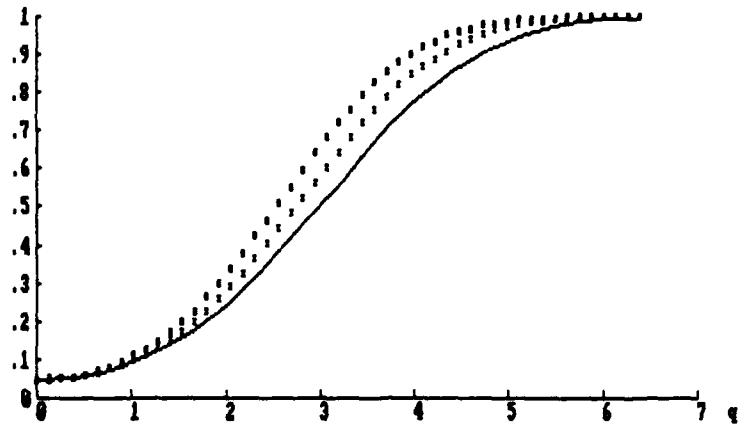

Fig. 20. *Power of the chi-square, subset chi-square and the independent tests method for alternatives in the direction of* $(1,1,1,1)$. *The blocks are the chi-square test, the* x*'s are the subset chi-square test and the solid line is the independent tests method.*

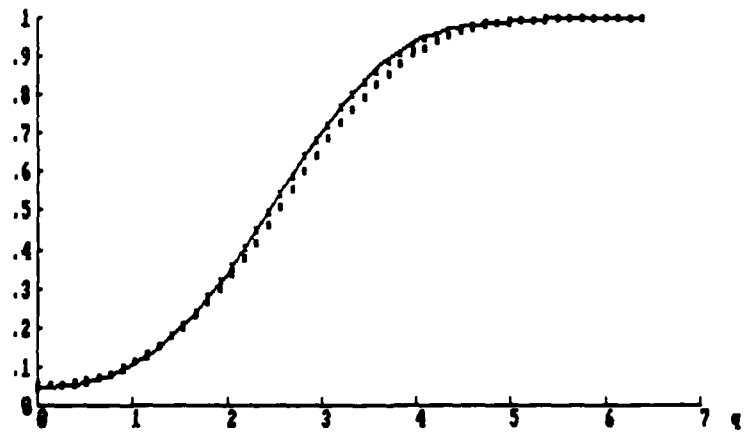

Fig. 21. *Power of the chi-square, subset chi-square and the independent tests method for alternatives in the direction of* $(1,0,0,0)$. *The blocks are the chi-square test, the* x*'s are the subset chi-square test and the solid line is the independent tests method.*

closer together than in Figure 20. The subset chi-square and the independent tests method are virtually identical and the standard chi-square is only slightly worse.

The subset chi-square represents a good compromise between the independent tests method and the chi-square test. It also possesses other features to recommend it. *How a test fits into a graphical, model selection environment* must also be considered. For instance, the chi-square test does not indicate which components caused the rejection, only that it did reject. The independent tests method does indicate which components are large but considers them only singly. Heuristically, it seems possible that S_1^2 and S_2^2 may be insignificant taken singly but that $S_1^2 + S_2^2$ may be significant. From practical experience, the independent tests method seems to be 'stingy' in declaring components significant. This will be seen, for example, in a data set examined in Section 5. The subset chi-square test does not suffer from these difficulties.

The subset chi-square test lends itself to graphical display much as AIC or CAT do. Define

$$C(k) = \max_{(i_1,\ldots,i_k)} S_{i_1}^2 + \cdots + S_{i_k}^2 - \chi_k^2(t(M,\alpha)),$$

for $k = 1,\ldots,M$, where the indices, $(i_1,\ldots,i_k)$, range over all subsets of size k of $\{1,\ldots,M\}$. One then graphs $C(k)$ versus k. If the values are all negative, the null hypothesis is not rejected. Since $C(k)$ is a function, it has shape and shapes impart information. If $C(k)$ has a very sharp and pronounced peak, then there is a strong choice for a particular subset. If $C(k)$ is flat without much of a peak, then there are several subsets which could be considered candidates. These interpretations will take on greater meaning in the next subsection where *an orthogonal series estimator* based on the components is introduced.

A plot of $C(k)$ versus k is only plotting the winner for each subset size. One could also plot below $C(k)$ the next largest value of the criterion function. This would give some indication of the cost of switching the smallest component in the optimal subset with the next smaller component. Since all these concepts are easily written in terms of order statistics, computation is not a problem.

There is no ready analytic method for determining the function $t(M,\alpha)$. Even for $M = 2$, the integral is not tractable. The function is found by simulation methods. The procedure is to fix $u = t(M,\alpha)$, the nominal size of the chi-squares, at various levels and the estimate the true size of the test, α. For the j^{th} realization of M iid N(0,1) random variables, the function

$$B_j(M,u_i) = \begin{cases} 1 & \text{if the test accepts for } t = u_i, \\ 0 & \text{if the test rejects for } t = u_i, \end{cases}$$

is found for the grid of u_i's equal to $u_i = 0.95 + 0.04999999(i-1)/25$ for $i = 1,\ldots,25$. These functions are then averaged over 25,000 replications to arrive at the function $\hat{R}(M,u)$:

$$\hat{R}(M,u_i) = \frac{1}{25{,}000}\sum_{j=1}^{25{,}000} B_j(M,u_i).$$

The function $\hat{R}(M,u)$ is estimating $R(M,u)$ which is the inverse function of $t(M,\alpha)$. Multiple uses are made of the M random variables for each realization since the test is conducted at a grid of u values. However, this preserves the monotonicity of the estimated function $\hat{R}(M,u)$. The value of $t(M,\alpha)$ is found by interpolating the function $(\hat{R}(M,u_i),u_i)$. That is, one finds i and i' with $i = i' - 1$ such that

$$\hat{R}(M,u_i) \leq 1-\alpha$$
$$\hat{R}(M,u_{i'}) \geq 1-\alpha$$

and then linearly interpolates the pairs $(\hat{R}(M,u_i),u_i)$, $(\hat{R}(M,u_{i'}),u_{i'})$ at the point $1-\alpha$ to arrive at $\hat{t}(M,\alpha)$ calculated from the u domain. From this procedure, it is clear why it is so important that the estimated function $\hat{R}(M,u)$ be monotone. If it were not, an inverse wouldn't exist and the procedure for estimating $t(M,\alpha)$ would fail.

Figure 22 presents the function $(u,\hat{R}(M,u))$ for $M = 2$. The function has been linearly interpolated between the points u_i. The function is indeed monotone as needed. The simulations are conducted for $M = 2,\ldots,15$. In this framework, presenting 14 graphs would be somewhat awkward. Instead, Table 5

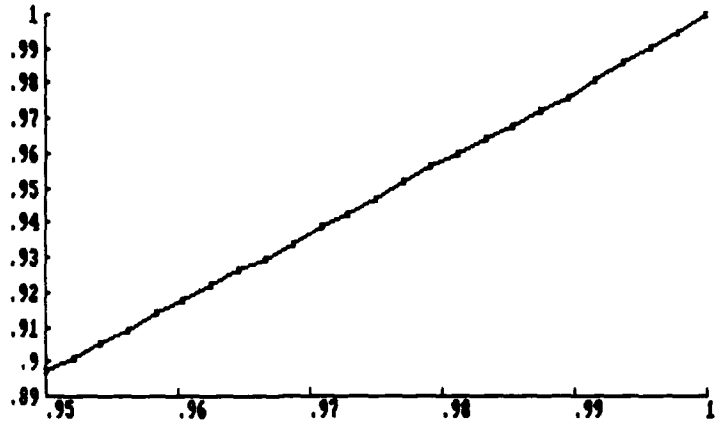

Fig. 22. *The function $\hat{R}(M,u)$ for determining the critical sequence of the subset chi-square test for $M = 2$.*

presents $\hat{t}(M,\alpha)$ for $\alpha = 0.05$ and 0.01 and for $M = 2,\ldots,15$. These values can be used in conjunction with an algorithm to evaluate the chi-square quantile to find the critical sequence.

In this subsection a new test, the subset chi-square test, was introduced and compared to two existing tests, the chi-square and the independent tests method. The subset chi-square was seen to be a good compromise between these other two in terms of the kinds of alternatives it detects. Further, the subset chi-square lends itself to the kind of graphical, model selection techniques which are sought. It is possible to define a criteria function, $C(k)$, which not only indicates acceptance and rejection but also points out particular subsets which are deemed significant. A simulation study was conducted to estimate the function $t(M,\alpha)$ which determines the sequence of critical values for the subset chi-square test. In the next subsection, the subset chi-square test is applied to the components. In this case, important uses are made of the components found to be significant.

3.3.6. *Orthogonal Series Estimates.* In this subsection, the subset chi-square test is applied to the components. This test leads naturally to an orthogonal series estimator of the comparison density. The relation of the orthogonal series

Table 5

Points at which to evaluate the chi-square quantile to find the critical sequence for the subset chi-square test (values are multiplied by 1000).

	Size of Overall Test	
M	0.05	0.01
2	976.29	995.44
3	985.33	997.36
4	990.08	998.30
5	992.44	998.66
6	994.19	998.94
7	995.82	999.22
8	996.51	999.37
9	997.29	999.49
10	997.77	999.56
11	998.22	999.64
12	998.51	999.70
13	998.64	999.73
14	998.86	999.77
15	998.93	999.79

estimator and the boundary kernel estimator is investigated. The orthogonal series estimator is found to be a damped series with the weights being the eigenvalues of the null covariance kernel. Hence, one can view the boundary kernel as behaving like a damped orthogonal series estimator.

The subset chi-square test can be applied directly to the sample normalized components, $Z_{N1}, \ldots, Z_{NM}$. This application is justified by Lemma 3.3.3 which gives the joint convergence in distribution of these random variables to limiting random variables which are iid N(0,1) under H_o. The first result of the application of the subset chi-square test to the components is invariance.

Lemma 3.3.5. *The subset chi-square test applied to the components is invariant as to which sample is called the first.*

The intriguing idea of the subset chi-square test is that it returns subsets of significant components as well as an accept/reject decision. Since these compo-

nents are also interpretable as generalized Fourier coefficients, it is only natural to define an orthogonal series estimator based on these. Suppose $C(k)$ attains its maximum at $k = k^*$ and that $C(k^*) > 0$. Let $i_1, \ldots, i_{k^*}$ be the subset which generates this maximum value. An orthogonal series estimate of $d_{(N)}$ is

$$\hat{d}_{h,M}(w) - 1 = \sum_{j \in (i_1, \ldots, i_{k^*})} Z^*_{Nj} \phi^h_j(w).$$

Other subsets could be examined based on the shape of $C(k)$.

Before examining $\hat{d}_{h,M}(w)$ any further, the relation of $\hat{d}_h$ and

$$\hat{d}_{h,\infty}(w) - 1 = \sum_{j=1}^{\infty} Z^*_{Nj} \phi^h_j(w)$$

should be looked into. As pointed out in Subsection 3.3.4, if $\{\phi^h_j\}$ are complete for S^h_D then equation (3.3.3); applies otherwise (3.3.5) holds. More than this can be said, however. In Subsection 2.3.4 the relation between Fourier based estimates and kernel estimates was pointed out. This relation is a two way street. The best way to examine $\hat{d}_{h,M}$ or $\hat{d}_{h,\infty}$ is not as a decomposition of $\hat{d}_h(w) - 1$ but as a function of the original data. Recall the representation of Z^*_{Nj} as

$$\begin{aligned} Z^*_{NJ} &= \int_0^1 J^h_j(u) dD_N(u) - \int_0^1 J^h_j(u) du \\ &= \int_0^1 J^h_j(u) d[D_N(u) - u] \\ &= \int_0^1 \left[J^h_j(u) - \int_0^1 J^h_j(t) dt \right] dD_N(u), \end{aligned} \tag{3.3.8}$$

where J^h_j is defined in Subsection 3.3.3. The large sample mean is used here instead of the small sample mean to avoid a dependence on N.

It is very interesting to note that the family of score functions, $\{J^h_j(u)\}$, is not orthogonal. The components cannot be regarded as the Fourier coefficients of an orthogonal decomposition of $D_N(u) - u$. Instead, they satisfy the condition

$$\int_0^1 J^h_j(u) J^h_{j'}(u) du - \int_0^1 J^h_j(w) dw \cdot \int_0^1 J^h_{j'}(t) dt = 0,$$

for $j \neq j'$. It does follow from this condition that the sequence $\{O^h_j(u)\}$ is orthogonal where $O^h_j(u) = J^h_j(u) - \int_0^1 J^h_j(t) dt$. Thus, the components Z^*_{Nj} can

be regarded as the generalized Fourier coefficients resulting from the usual orthogonal series density estimation formulas using the orthogonal functions $O_j^h(u)$ and applied to the normalized ranks, $R_1/N, \ldots, R_m/N$. This observation is clear from the last equality of equations (3.3.8). Since each $O_j^h(w)$ integrates to zero, the constant function 1 is in this basis as well. Its Fourier coefficient is always 1 and is subtracted on the left hand side of the equals sign in the defining formulas for $\hat{d}_{h,M}$ and $\hat{d}_{h,\infty}$.

Notice it is claimed that the $O_j^h(u)$'s are orthogonal and not orthonormal. Figures 15 through 17 presented the score functions, J_j^h, for $j = 1,2,3,4$ and $h = 0.5, 0.3, 0.1$. At the time, the shapes of these score functions and not their magnitudes were of primary interest. Hence, each was normalized to attain a maximum absolute value of 1. Now, however, the magnitudes are of interest. Figures 23 through 25 present the orthogonal functions $O_j^h(u)$ for $j = 1,2,3,4$ and $h = 0.5, 0.3, 0.1$. Figure 23 is an overlay plot of the 4 functions for $h = 0.5$; Figure 24 an overlay for $h = 0.3$; and Figure 25 an overlay for $h = 0.1$. The most striking feature is that the magnitudes of the O_j^h's decrease with increasing j. The second striking feature is that this rate of decrease is slower with the smaller bandwidths. One suspects an interplay between this observation and the slower rate of decline in the eigenvalues (recall Figure 14) for decreasing h. The following steps make this relationship mathematically clear:

$$
\begin{aligned}
\|O_j^h\|^2 &= \int_0^1 O_j^h(u)^2 du \\
&= \int_0^1 \left[\int_0^1 K_s\left(\frac{w-u}{h}\right) \phi_j^h(w) dw - \int_0^1 \phi_j^h(t) dt \right]^2 du \\
&= \int_0^1 \left(\frac{1}{h} \int_0^1 \frac{1}{h} K_s\left(\frac{w-u}{h}\right) \phi_j^h(w) dw \cdot \int_0^1 K_{s'}\left(\frac{t-u}{h}\right) \phi_j^h(t) dt \right) du \\
&\quad - 2 \int_0^1 \phi_j^h(t) dt \cdot \int_0^1 \int_0^1 \frac{1}{h} K_s\left(\frac{w-u}{h}\right) \phi_j^h(w) dw du \\
&\quad + \left[\int_0^1 \phi_j^h(t) dt \right]^2
\end{aligned}
$$

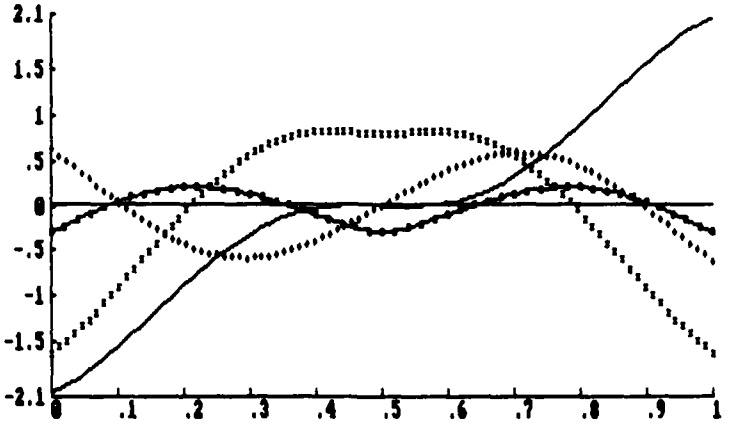

Fig. 23. *The orthogonal functions* $O_j^h(u)$ *for* $j = 1,2,3,4$ *and* $h = 0.5$. *The function* $O_1^{0.5}(u)$ *is the solid line;* $O_2^{0.5}$ *is the broken line of* x*'s;* $O_3^{0.5}$ *is the broken line of* $+$*'s;* $O_4^{0.5}$ *is the solid line with blocks. The square of the* $\mathcal{L}_2$ *norms of these functions are:* 1.0585, 0.6574, 0.1619, 0.0275.

$$
\begin{aligned}
&= \int_0^1 \int_0^1 \int_0^1 \frac{1}{h^2} K_s\left(\frac{w-u}{h}\right) K_{s'}\left(\frac{t-u}{h}\right) \phi_j^h(w)\phi_j^h(t)dwdtdu \\
&\quad - \left[\int_0^1 \phi_j^h(t)dt\right]^2 \\
&= \int_0^1 \int_0^1 \phi_j^h(w)\left[\int_0^1 \frac{1}{h^2} K_s\left(\frac{w-u}{h}\right) K_{s'}\left(\frac{t-u}{h}\right) du - 1\right] \phi_j^h(v)dwdv \\
&= \int_0^1 \int_0^1 \phi_j^h(v) C_h(v,w) \phi_j^h(w)dwdv \\
&= \theta_j^h,
\end{aligned}
$$

where $s = s(w,h)$ and $s' = s(t,h)$. Hence, the square of the $\mathcal{L}_2$ norm of the function O_j^h is equal to the eigenvalue θ_j^h.

This proves conclusively that $\hat{d}_{h,M}$ and $\hat{d}_{h,\infty}$ are damped orthogonal series estimators where the weights are the eigenvalues. By this is meant that the estimator has the following representation:

$$
\hat{d}_{h,\infty}(w) = 1 + \sum_{j=1}^{\infty} \theta_j^h \left(\frac{Z_{Nj}^*}{\sqrt{\theta_j^h}}\right)\left(\frac{O_j^h(w)}{\sqrt{\theta_j^h}}\right),
$$

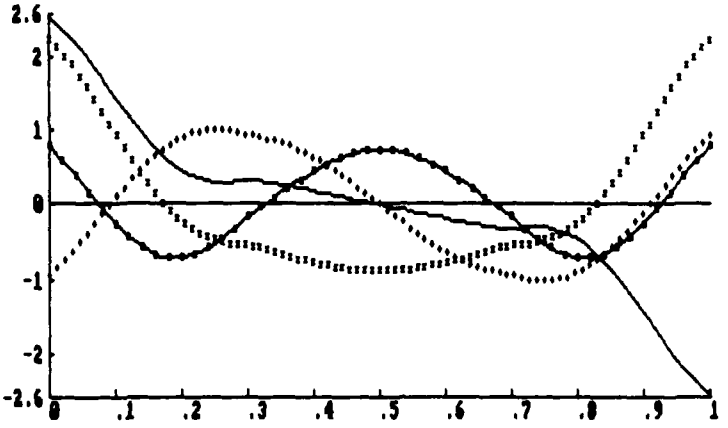

Fig. 24. *The orthogonal functions* $O_j^h(u)$ *for* $j = 1,2,3,4$ *and* $h = 0.3$. *The function* $O_1^{0.3}(u)$ *is the solid line;* $O_2^{0.3}$ *is the broken line of* x*'s;* $O_3^{0.3}$ *is the broken line of* $+$*'s;* $O_4^{0.3}$ *is the solid line with blocks. The square of the* $\mathcal{L}_2$ *norms of these functions are:* 1.0921, 0.9008, 0.4941, 0.2547.

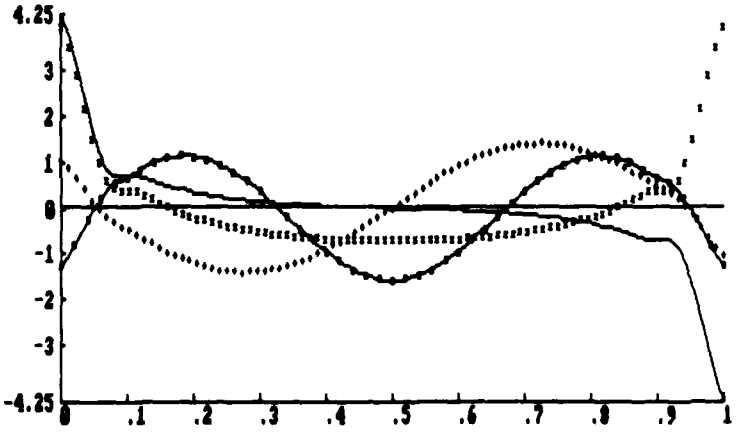

Fig. 25. *The orthogonal functions* $O_j^h(u)$ *for* $j = 1,2,3,4$ *and* $h = 0.1$. *The function* $O_1^{0.1}(u)$ *is the solid line;* $O_2^{0.1}$ *is the broken line of* x*'s;* $O_3^{0.1}$ *is the broken line of* $+$*'s;* $O_4^{0.1}$ *is the solid line with blocks. The square of the* $\mathcal{L}_2$ *norms of these functions are:* 1.1034, 1.0654, 0.9317, 0.8654.

since the sequence $\{O_j^h(u)/\sqrt{\theta_j^h}\}$ is now orthonormal. If the sequence $\{\phi_j^h\}$ is complete for S_D^h, one can regard the boundary kernel estimator $\hat{d}_h$ as being equivalent to a damped orthogonal series estimator. If the sequence is not complete for S_D^h, this interpretation is limited to the projection of $\hat{d}_h(w) - 1$ onto the subspace S_S^h.

The fact that the series is weighted makes the usual problem of the choice of truncation point much less difficult. The representation for $\hat{d}_{h,\infty}$ provides an alternate and intriguing explanation of the role of the bandwidth in determining the smoothness of the estimate. Larger bandwidths lead to smoother estimates than smaller bandwidths (recall the discussion of Subsection 2.4.3). From Figures 23 through 25 it is apparent that the higher order orthogonal functions are rougher (higher frequency) than the lower order ones. A smaller bandwidth gives more weight to the higher order O_j^h's than a larger bandwidth, hence the smaller bandwidth is capable of producing rougher estimates.

One can also appreciate why the estimate $\hat{d}_h$ is consistent as $h \to 0$. As the bandwidth shrinks, more and more of the basis is allowed to enter and contribute materially to the shape of the final estimate. In the limit, any shape can be duplicated. This also explains the increase in variance as h decreases. With decreasing bandwidths, the number of parameter estimates (components) that make up the estimate is increasing as more are given significant weight. This results in an increase in variance.

At this point, it may be wise to summarize the properties of $\hat{d}_h$ and $\hat{d}_{h,M}$. In fixed samples each estimator is biased; this is to be expected for any density estimator [see Rosenblatt (1956) or Seheult and Quesenberry (1971)]. For fixed h the estimators are also asymptotically biased. But this is not the whole story. As a practical matter, with increasing amounts of data one would be lead naturally to choosing smaller h and larger M. These actions reduce the amount of bias present. Indeed, such a process will even attain consistency as per Theorem 3.2.2. The distinction is between what one assumes to prove theorems and what one does in implementing a procedure. One uses fixed-h results such as Theorem 3.2.3 and Lemma 3.3.3 to find approximate distributions. One merely conceives

of the bandwidth being fixed as the sample size increases. Although $\lambda_{(N)}$ is not selectable, there is still a very strong analogy between the treatment of $\lambda_{(N)}$ and h. It is assumed that $\lambda_{(N)}$ converges to λ_0, $0 < \lambda_0 < 1$, as $m \wedge n \to \infty$. The limit stochastic processes all contain the term λ_0. Yet given just one sample, one can only conceptualize a convergence. For the single sample, convergence has no meaning.

It is not guaranteed that either $\hat{d}_h$ or $\hat{d}_{h,M}$ will be non-negative and integrate to 1. The estimates may themselves not be densities. If the density is not non-negative, then having it integrate to 1 is of little benefit. The decision to employ such estimators was made in Section 2. Suffice it to say here that one must recall what the estimate of $d_{(N)}$ is used for. This relates to the discussion of bias as well. The important aspect of $\hat{d}_h$ and $\hat{d}_{h,M}$ is their shape. It is not intended to use them as density estimates. One will not simulate random variables from them. Their important interpretation is that of likelihood ratio. These other properties would be nice, but are not at all essential.

More importantly, the orthogonal series estimator can be shown to satisfy invariance even in finite samples. This result is stated as Lemma 3.3.6.

Lemma 3.3.6. *The orthogonal series estimate obeys the invariance condition*

$$\lambda_{(N)} d^F_{h,M}(w) + (1 - \lambda_{(N)}) d^G_{h,M}(w) = 1$$

in finite samples.

The function $d^F_{h,M}(w)$ is the estimate when the population with distribution function F is called the first sample. The function $d^G_{h,M}(w)$ is the estimate when the population with distribution function G is called the first sample. This result is due to the small sample mean correction to the components which caused them to be invariant.

It is now known also that $\hat{d}_{h,M}$ is a damped orthogonal series estimate and that the weights are the eigenvalues. Only a subset of the components (or frequencies) making up $\hat{d}_h(w)$ are present in $\hat{d}_{h,M}$. The estimate of $\hat{d}_{h,M}$ can be smoother than $\hat{d}_h$ but not rougher in the sense higher frequencies may be absent

from $\hat{d}_{h,M}$. This observation will lead to a suggestion for choosing the bandwidth in the next subsection.

In this subsection, the subset chi-square test was applied to the normalized components, $Z_{N1},\ldots,Z_{NM}$. It was seen that this test leads naturally to an orthogonal series estimator, $\hat{d}_{h,M}$. A representation for this estimator in terms of the original data was found. It was seen to be a damped orthogonal series estimator with the weights being equal to the eigenvalues of the null covariance kernel. The properties of the estimators $\hat{d}_h$ and $\hat{d}_{h,M}$ were discussed. They were seen to be biased, even asymptotically-so under a fixed bandwidth regime. However, it was argued that such a statement is vacuous in the sense that the bandwidth will naturally decline with increasing sample size and that fixed bandwidth theorems are useful for approximating distributions. One is not meant to seriously consider keeping the same bandwidth out to infinite sample sizes. The goal of any density estimation technique should be to select the bandwidth to fit the data parsimoniously whatever the resulting bandwidth might be.

3.3.7. *Choosing h and M*. In this subsection, several schemes for choosing h and M are examined. One can choose h either graphically or by an automatic criterion. One can choose M to cover only the desired alternatives or to include all eigenvalues above some cutoff. The issues involved in choosing h from the data on the properties of the test are also discussed.

The usual density estimator has either h or M; here both are present. This adds flexibility to the problem and is not a hindrance. There are several philosophies that might be adopted. The first is based on a remark from Subsection 3.3.6 that the orthogonal series estimate can be smoother than the boundary kernel estimate but not rougher. This approach would suggest choosing h to undersmooth the data (*i.e.* $\hat{d}_h$ is slightly too rough) and then choose M to include all the components whose eigenvalues exceed some cutoff such as 0.01 or 0.001. One then relies on the subset chi-square test to include or exclude the components as appropriate. Thus, the orthogonal series estimate of $d_{(N)}$ has available to it all models from too smooth to too rough. Table 6 gives the numbers of eigenvalues above three cutoffs as a function of bandwidth.

Table 6

Number of eigenvalues of the null covariance kernel above a cutoff.

	Cutoff		
h	0.01	0.001	0.0001
0.5	4	5	6
0.4	5	6	8
0.3	6	8	10
0.2	9	12	16
0.1	16	23	32

Another alternative is to use a criterion function such as least squares cross-validation (LSCV) to choose the bandwidth and then to include all the components above some cutoff. The properties of LSCV have not been established in this setting. Anyone proceeding upon such a path should use caution. These first two suggestions are similar in spirit.

A completely different approach is to fix M. One might fix M based on the types of alternatives one is considering, for example, $M = 2$ for location and scale. Having fixed M one is free to choose h. One could choose h based on fit or based on the types of distributions one wishes to best protect against, that is, one could choose h so that the shapes of the score functions are pleasing. Overall, there seem to be very good opportunities to direct the procedure toward more specific alternatives if this type of information is available.

The first two procedures and possibly the third involve the selection of the bandwidth based on the data. The bandwidth in these cases is not only random but also a function of the data. It is of interest how the properties of the test might be affected. This sort of problem is not at all unheard of in statistics. An analogue in regression would be the distribution of the parameter t-statistics under a regression selection criterion like stepwise regression. If the bandwidth were random but not a function of the data, then the answer would be trivial: the size of the test would be unaffected. Of course, the bandwidth will always depend on the data and the situation is more complicated. The effect of a data-driven bandwidth depends on exactly how the bandwidth depends on the

data. For graphical selection procedures, this question is not resolvable because it cannot be quantified. For criteria functions such as LSCV there is hope for an answer though it most probably would result from simulation rather than analytic techniques. An adjustment for a data-driven bandwidth would amount to adjusting α in the critical sequence defining the subset chi-square test.

These simulations are not performed here. Instead, they are left as future work. There are cases where completely automatic methods may be appropriate; say, when the output density is required as input to another procedure. Otherwise, those using an automatic method who do not check the fit of the estimated density may be rudely surprised as such methods do fail: LSCV is known, for example, to drastically undersmooth about 5–20% of the time [see Hart (1988)]. These methods are more properly used to suggest choices of bandwidths. It is up to the user of these techniques to make the final choice based on other criteria such as fit.

One of the strong points of nonparametric density estimation in general is its ability to suggest different models. The methodology here is no different. Aside from altering bandwidths, the function $C(k)$ is capable of suggesting quite different shapes for a given bandwidth. If the function is nearly level, then several quite different models may result. However, caution should be made against one particular abuse. Sometimes it will occur that the subset chi-square will fail to reject. Upon examining the components, one may see that by decreasing M or changing h that the test would reject. It is statistically dishonest to make such a modification and declare significance. Such a procedure can drastically alter the properties of the test. Since some choice of M and h must be made, this choice should be made before the subset chi-square test is run. These procedures will still affect the properties of the test, but they will do so in a much less egregious manner.

In this subsection, several different methods of choosing the bandwidth and the truncation point were examined. Which to use is determined by the objectives of the researcher. The issues involved in the effect of a data-driven bandwidth on the size of the subset chi-square test were discussed. In order to adjust the size

of the test, one would have to pick a specific, quantifiable choice criterion. The necessary adjustment would most probably have to be determined by simulation techniques.

3.3.8. *Summary of the Unified Testing and Estimation Procedure.* The pieces of the unified testing and estimation procedure have been scattered throughout this section. They are brought together in this subsection. The procedure is seen to fulfill the features outlined in Subsection 2.2.6 and Section 1. This procedure is summarized by the following list:

1. Univariate Analysis
2. Preliminary Two Sample Analysis
3. Choosing M and h
4. Executing the Subset Chi-Square Test on the Components
5. Plotting the Orthogonal Series Estimate of the Comparison Density

Any two sample analysis should start with three univariate analyses: the two individual samples and the pooled sample. Statistics such as the mean, median, standard deviation, twice the interquartile range and trimmed means should be examined. Identification quantile plots [Parzen (1979)] should be constructed. The philosophy is that before asking if F and G are equal, it is best to investigate the properties of each on their own. Examining the pooled sample can highlight distinctions between the two.

During the Preliminary Two Sample Analysis stage, visual indications of the fit of the two samples and standard two sample statistics are given. An overlay plot of the two identification quantile plots is given as is an QQ plot. These two plots remove the effect of location and scale: they compare the shapes of the distributions. Traditional statistics such as the Cramér-von Mises or Anderson-Darling are also given at this stage.

At this point, M and h need to be chosen by one of the methods outlined in Subsection 3.3.7. In the example in Section 5, they will be chosen by the first method outlined. The components are then calculated and the subset chi-square test applied to them. The criteria function, $C(k)$, as defined in Subsection 3.3.5

should be displayed. The graph should include a horizontal reference line at 0. If $C(k)$ does not exceed zero, the null hypothesis cannot be rejected and the estimate of the comparison density is the uniform density. If $C(k)$ exceeds 0 at one or more k values then estimates of the comparison density based on the eigenfunctions should be displayed. The estimate based on the subset which maximizes $C(k)$ is always displayed. Others may be displayed at the user's discretion based on the shape of $C(k)$. Along with these graphs a list of which components are significant and the components themselves should be given. Each graph should include a horizontal reference line at 1. The boundary kernel estimate can also be overlaid for reference.

Note that this procedure does indeed fulfill the criteria outlined in Subsection 2.2.6 and Section 1. It is certainly a graphically oriented technique. The subset chi-square test applied to the components is also a selection procedure for a model of $d_{(N)}(w)$. If the null hypothesis cannot be rejected the model is uniformity; if the null hypothesis is rejected the model is the orthogonal series estimate corresponding to those components found significant. The test is omnibus. In fact, the breadth of the class protected against is under the control of the user. The distribution of the components is nonparametric distribution free under H_o since they are linear rank statistics. There are as few restrictions placed on F and G as possible while maintaining weak convergence results for the comparison distribution empirical process. Finally, the estimation of the relation of F to G is given by the estimate of the comparison density. All the requirements are fulfilled by this methodology.

In summary, this section has detailed the theoretical and computational aspects of the boundary kernel estimate of the comparison density and tests of its uniformity. The section started by giving pointwise results for the boundary kernel estimator under a bandwidth shrinking to zero, the pointwise asymptotic normality of the boundary kernel estimator under H_o; its pointwise consistency and invariance under general alternatives was shown. A stochastic process called the kernel density process was defined from the boundary kernel estimator. Conditions were given for its weak convergence to a limiting process under a fixed

bandwidth. A rationale for fixing the bandwidth was given.

Tests of the null hypothesis were based on the kernel density process. It was argued that the best strategy was to base any test on a fixed number M of the components of the kernel density process. The components were defined as the inner product of the eigenfunctions of the covariance kernel of the kernel density process under H_o and the boundary kernel estimate less 1, $\hat{d}_h(w) - 1$. Properly scaled, these components converge jointly to iid N(0,1) random variables under H_o. The components are interpretable both as generalized Fourier coefficients and as rank statistics.

A new test, the subset chi-square test, was introduced and compared to existing tests. This test was then applied to the components. The subset chi-square test was seen to have several desirable properties. First, it considers the components in combination not just singly. Second, it indicates which components are deemed large. Third, it lends itself to graphical display. The subset chi-square test was also seen to suggest an orthogonal series estimate of the comparison density based on the components and the eigenfunctions. The relation between the orthogonal series estimate and the boundary kernel estimate was explored.

Methods of choosing the bandwidth and truncation point were examined. The implications of data based choices of the bandwidth were also discussed. Finally, the methodology was summarized. It was seen to truly meet the criteria outlined in Section 1 and Subsection 2.2.6.

4. POWER AND SIZE STUDIES

4.1. Introduction

In this section, power and size studies are conducted. Subsection 4.2 covers power studies; subsection 4.3 covers size studies. Subsection 4.2 derives and explains the theoretical concepts necessary for defining power functions and the asymptotic relative efficiency between two rank statistics. Also detailed are the simulation and numeric techniques used to actually find the power functions.

The asymptotic relative efficiencies of the first two components are compared to standard rank tests. The bandwidth is seen to have an effect on this efficiency. The first component is generally less efficient than the standard rank statistics and the second more efficient. It is found that the optimal choice of bandwidth is not necessarily the same for both location and scale alternatives of the same underlying distribution. A good compromise choice of bandwidth, however, can be made for the distributions considered.

The subset chi-square test applied to the components of the kernel density process is found to have very good power properties. The Cramér-von Mises and Anderson-Darling statistics have good power against the location alternatives examined. However, when the alternative starts to principally affect higher components, these two statistics have much poorer power properties. The subset chi-square test is equally good against any alternative which affects components it considers. It outperforms the Cramér-von Mises and Anderson-Darling statistics by wide margins for alternatives influenced mainly by the fourth and higher components.

It is seen that the key to the power of the subset chi-square is the choice of truncation point (M) and bandwidth (h). Choosing the truncation point too large (h too small) reduces power because the signal is swamped by noise. Choosing the truncation point too small (h too large) reduces power because the signal is missed by the test. Careful selection of the truncation point and bandwidth should chart a course between these twin abysses.

Subsection 4.3 covers the small sample size of the subset chi-square test applied to the components. The size remains very close to the appropriate value even for samples as small as $n = m = 5$. The size deviates significantly when one of two situations occurs. The first occurs when the bandwidth is clearly chosen to be too small. For instance, the size deviates substantially from its nominal value if $h = 0.3$ or $h = .2$ is used for $n = m = 5$. One would never choose such a small bandwidth for this sample size in practice. The second occurs when m is very small (say $m = 5$) and $n \gg m$ (say $n = 100$). Again, one doesn't expect to see such cases very often in practice.

4.2. Power Studies

4.2.1. *Introduction.* Subsection 4.2 investigates the power of the procedures discussed in Section 3 along with the the Cramér-von Mises and Anderson-Darling statistics. The power functions in this subsection are asymptotic. Subsection 4.2.2 discusses the notion of a local alternative. The main result of this subsection is a theorem stating the conditions under which $\mathrm{CDo}_N = \sqrt{N}[D_N(u) - u]$ still converges weakly under local alternatives. Local location, scale and Fourier alternatives are defined. The asymptotic relative efficiency of two rank statistics is defined.

Subsection 4.2.3 gives the methods to be used to find power functions. A method based on simulation is described for the subset chi-square test. A method to numerically invert the characteristic function is also described. A theorem about its numerical consistency is proved. This method is used to find the power functions for the φ^2_h, Cramér-von Mises, and Anderson-Darling statistics.

Subsection 4.2.4 demonstrates the calculations necessary to check the conditions for weak convergence of CDo_N. These conditions are checked for Cauchy location and scale alternatives. Subsection 4.2.5 presents power curves for two distributions for both location and scale alternatives. Curves for two Fourier alternatives are also given. The subset chi-square test applied to the components is seen to perform credibly, particularly for alternatives stressing the second and higher components. Asymptotic relative efficiencies comparing the first two

components to standard rank tests for four underlying distributions are given. The bandwidth is seen to make a difference. No statistic dominates over all the different distributions.

4.2.2. *Weak Convergence of* CDo_N *Under Local Alternatives.* The weak convergence of the stochastic process CDo_N to a limiting process $L + (1-\lambda_0)^{-1/2}\Delta$ under local alternatives is proved in this subsection. This is very important since under fixed alternatives one expects the process to become unbounded as the sample sizes increase. From the results of Section 2, one can only claim that CDo_N converges weakly under H_o, because one has the identity $\mathrm{CDo}_N(u) \equiv \mathrm{CD}_N(u)$. It is the goal of this subsection to broaden the class for which convergence is claimed to include local alternatives.

The discussion opens by examining the concept of a local alternative. The statement of the theorem on weak convergence follows. Types of local alternatives satisfying the conditions are discussed. Local location and scale alternatives are defined. Also introduced is the strategy of defining the local alternative by parametrically defining its limiting bias function and not specifying the sequence of underlying distributions.

A local alternative is an alternative in which the distribution of $Y_1, \ldots, Y_n$ depends on n so one has $Y_{n1}, \ldots, Y_{nn}$ are iid according to the distribution function $G_{(n)}$. The problem is that for fixed $G \neq F$, a test is either consistent or inconsistent. If it is consistent, the asymptotic power function is 1, if it is inconsistent the power is equal to the size of the test. Hence, asymptotic power curves drawn for fixed alternatives are very uninteresting. Instead, one chooses a sequence $\{G_{(n)}\}$ of alternatives such that $G_{(n)} \to F$ as $n \to \infty$. If $G_{(n)}$ converges to F at the correct rate, several good things happen. First, the power functions are not degenerate, that is, they are not uniformly equal to α or 1. Second, statistics have the same distribution as under H_o with the exception of the addition of a non-zero mean. Typically, this makes power curves much easier to construct.

For the purposes of local alternatives, let

$$Y_{n1},\ldots,Y_{nn} \text{ be iid } G_{(n)},$$
$$H_{(N)}(x) = \lambda_{(N)}F(x) + (1-\lambda_{(N)})G_{(n)},$$
$$Q^{H_{(N)}}(u) = H^{-1}_{(N)}(u),$$

and

$$D_{(N)}(u) = FQ^{H_{(N)}}(u).$$

It is assumed that $G_{(n)}(x) \to F(x)$ at a rate so that the limit function, $\Delta(u)$, defined by

$$\Delta(u) = \lim_{m\wedge n\to\infty} \sqrt{n}[D_{(N)}(u) - u] \tag{4.2.1}$$

exists and is continuous with $\Delta(0) = \Delta(1) = 0$. Although not specified in the notation, $\Delta(u)$ will depend on λ_0 as well as a parameter γ which indexes the local alternative. The function $\Delta(u)$ is the bias function and Theorem 4.2.1 gives the conditions under which $\mathrm{CDo}_N \Rightarrow L + (1-\lambda_0)^{-1/2}\Delta$.

Theorem 4.2.1. *Assume that there exists $\Delta(u)$ such that $\|\sqrt{n}[D_{(N)}(u) - u] - \Delta(u)\| \to 0$ as $m\wedge n \to \infty$ and suppose there exists sequences of constants $\{a_{(N)}\}$ and $\{b_{(N)}\}$ such that*

$$\mathrm{P}[a_{(N)} \le Q^H_N(t) \le b_{(N)}, 0 \le t \le 1] \to 1$$

as $m\wedge n \to \infty$. Let $e_{(N)}$ and $f_{(N)}$ be defined as

$$e_{(N)} = Q^{H_{(N)}}(1/N) \wedge a_{(N)},$$
$$f_{(N)} = Q^{H_{(N)}}(1-1/N) \vee b_{(N)},$$

and suppose that

$$\sup_{e_{(N)}\le x\le f_{(N)}} \frac{f(x)}{g_{(n)}(x)} \to 1$$
$$\inf_{e_{(N)}\le x\le f_{(N)}} \frac{f(x)}{g_{(n)}(x)} \to 1$$

as $m \wedge n \to \infty$. *Then*

$$\mathrm{CDo}_N \Rightarrow L + (1-\lambda_0)^{-1/2}\Delta$$

in $(C[0,1], C_\rho, \rho)$ *as* $m \wedge n \to \infty$ *where* $L(u) = \sqrt{\frac{1-\lambda_0}{\lambda_0}}B(u)$ *and* $B(u)$ *is a Brownian bridge process.*

The norm $\|\cdot\|$ is the sup-norm. This theorem is the basic building block for deriving power functions. It follows from a proof entirely analogous to that of Theorem 3.2.3 that under the same conditions as Theorem 3.2.3

$$\mathrm{KDPo}_{N,h} \Rightarrow \mathrm{KDPo}_h + \delta_h,$$

where

$$\delta_h(w) = (1-\lambda_0)^{-1/2}\int_0^1 \frac{1}{h^2}K_s'\left(\frac{w-u}{h}\right)\Delta(u)du,$$

and KDPo_h is the process KDP_h when H_o is true.

To this point, nothing has been said about the character of $G_{(n)}$ or the rate at which it must converge to F for a non-degenerate limit function to exist. These details are now given. For a local location alternative, define $G_{(n)}$ by

$$G_{(n)}(x) = F(x - \gamma/\sqrt{n}).$$

Prihoda (1981) shows that the limit function is

$$\Delta(u) = (1-\lambda_0)\gamma f Q^F(u), \tag{4.2.2}$$

although a proof of this is embedded in the proof of Lemma 4.2.1 (below) as well. Pointwise convergence is easily shown, however Theorem 4.2.1 calls for uniform convergence since the sup-norm is used. Lemma 4.2.1 gives the conditions on F under which pointwise convergence implies uniform convergence for local location alternatives.

For a local scale alternative, $G_{(n)}$ is defined by

$$G_{(n)}(x) = F\left(\frac{x}{1+\gamma/\sqrt{n}}\right).$$

Prihoda (1981) shows the limit function in this case to be

$$\Delta(u) = (1 - \lambda_0)\gamma Q^F(u) f Q^F(u). \tag{4.2.3}$$

Under certain conditions on F, the convergence for location and scale alternatives can be shown to be uniform. Lemma 4.2.1 gives the result.

Lemma 4.2.1. *Let $\Delta(u)$ be given by (4.2.2) for a local location alternative and by (4.2.3) for a local scale alternative and suppose $\Delta(u)$ is continuous and $\Delta(0) = \Delta(1) = 0$. Suppose that f' exists and is bounded. Assume also that $\lambda_{(N)} = \lambda_0$. Then*

$$\|\sqrt{n}[D_{(N)}(u) - u] - \Delta(u)\| \to 0$$

as $m \wedge n \to \infty$.

These conditions are easily satisfied by the distributions to be considered. The remaining conditions of Theorem 4.2.1 must be shown on a case by case basis. These conditions are easier to show than uniform convergence.

There is another possible method for defining local alternatives. This is to ignore the underlying distributions F and $G_{(n)}$ and to define the limiting bias function $\Delta(u)$ parametrically. A convenient representation is

$$\Delta(u) = (1 - \lambda_0)\gamma \sum_{j=1}^{k} a_j \sin \pi j u. \tag{4.2.4}$$

The function $\Delta(u)$ preserves its known properties: $\Delta(u)$ is continuous with $\Delta(0) = \Delta(1) = 0$. This procedure is attractive for creating alternatives other than location and scale. As might be anticipated from the discussion in Section 3, location and scale alternatives affect mainly the first two components. Defining alternatives in this way allows one to easily put weight on the higher order components. Alternatives constructed in this way will be called Fourier alternatives since a sine basis is used to define $\Delta(u)$.

At this point, it is now possible to define the asymptotic relative efficiency (ARE) of two rank statistics. Let R_{N1} be a rank statistic with score function

$J_1(u)$ and let R_{N2} be a rank statistic with score function $J_2(u)$. Assume $J_1(u)$ and $J_2(u)$ are differentiable, then

$$\begin{aligned} R_{Ni} &= -\int_0^1 J_i'(u)\mathrm{CDo}_N(u) \\ &\xrightarrow{\mathrm{d}} -\int_0^1 J_i' L(u)du - (1-\lambda_0)^{-1/2}\int_0^1 J_i'(u)\Delta(u)du \end{aligned}$$

as $m \wedge n \to \infty$ for $i = 1, 2$, where the convergence in distribution follows from the weak convergence of CDo_N to $L + (1-\lambda_0)^{-1/2}\Delta$. Restating this result, one has

$$\frac{R_{Ni} + (1-\lambda_0)^{-1/2}\int_0^1 J_i'(u)\Delta(u)du}{\sqrt{\frac{1-\lambda_0}{\lambda_0}\left(\int_0^1 J_i(u)^2du - \left[\int_0^1 J_i(u)du\right]^2\right)}} \xrightarrow{\mathrm{d}} \mathrm{N}(0,1),$$

as $m \wedge n \to \infty$. The conditions of Noether's theorem [see Randles and Wolfe (1979), page 147] which justify defining ARE's are clearly met. The asymptotic relative efficiency of R_{N1} to R_{N2}, denoted $\mathrm{ARE}(R_{N1}, R_{N2})$, is defined as

$$\mathrm{ARE}(R_{N1}, R_{N2}) = \frac{\kappa_1^2}{\kappa_2^2},$$

where

$$\kappa_i = -\frac{(1-\lambda_0)^{-1/2}\int_0^1 J_i'(u)\Delta(u)du}{\sqrt{\frac{1-\lambda_0}{\lambda_0}\left(\int_0^1 J_i(u)^2du - \left[\int_0^1 J_i(u)du\right]^2\right)}},\ i = 1, 2,$$

and κ_i is known as the efficacy of the rank test. If $\mathrm{ARE}(R_{N1}, R_{N2}) > 1$, then R_{N1} is asymptotically relatively efficient compared to R_{N2}.

The motivation for these definitions is straightforward once one realizes that the asymptotic power function for R_{Ni} is calculated as

$$\begin{aligned} \beta(\gamma) &= \mathrm{P}_\gamma[\text{Reject } \mathrm{H}_o] \\ &= \lim_{m\wedge n\to\infty} \mathrm{P}[|R_{Ni}/\sigma_i| > z_{\alpha/2}] \\ &= \lim_{m\wedge n\to\infty} 1 - \mathrm{P}[-z_{\alpha/2} - \kappa_i \le Z_{Ni} - \kappa_i \le z_{\alpha/2} - \kappa_i] \\ &= 1 - [\Phi(z_{\alpha/2} - \kappa_i) - \Phi(-z_{\alpha/2} - \kappa_i)], \end{aligned}$$

where σ_i is the denominator in the definition of κ_i and $Z_{Ni} = R_{Ni}/\sigma_i$. Increasing κ_i causes Φ to be evaluated further in the tail which reduces the quantity in the brackets and increases the power function. The efficacy, κ_i, does provide an ordering of the powers and is meaningful.

Power functions and ARE's are given in Subsection 4.2.5 for local scale and location alternatives corresponding to four underlying distributions: normal, logistic, Laplace, and Cauchy. Power functions are also calculated for two Fourier alternatives.

Summarizing, in this subsection the concept of local alternative was defined. A theorem giving the weak convergence of the comparison distribution empirical process was given. Local location and scale alternatives were defined and lemmas concerning the uniform convergence of the biases were proved. Local Fourier alternatives were defined. The ARE of two rank statistics was defined and the relevance of the measure illustrated. In Subsection 4.2.3, techniques for actually calculating the power curves are given.

4.2.3. *Computing Power Curves.* Power curves are constructed for the subset chi-square test applied to the components of the kernel density process, φ_h^2, CVM (Cramér-von Mises statistic), and AD (Anderson-Darling statistic). Two separate techniques are used. For the subset chi-square procedure, simulation methods are used. For the others, the characteristic function is numerically inverted.

Since finding the percentage points of the subset chi-square test under H_o required simulation techniques, it is not at all surprising that finding the power function does as well. The technique is as follows. The parameters M and h are given. The asymptotic bias for each of the M components, Z_{Ni}^*, is

$$b_j^* = \int_0^1 \phi_j^h(w)\delta_h(w)dw,$$

where δ_h is as defined in Subsection 4.2.2. The normalized components have bias

$$b_j = \sqrt{\frac{\lambda_0}{1-\lambda_0}} b_j^* / \sqrt{\theta_j^h}.$$

Although the notation does not reveal it explicitly, Δ, δ_h and thus b_j all depend on the parameter γ which indexes the local alternative [see equations (4.2.2), (4.2.3), and (4.2.4)]. To find the power function, $\beta(\gamma)$, one needs to find the probability of the subset chi-square test rejecting H_o when given M independent normal random variables with variance 1 and means b_j, $j = 1, \ldots, M$. The beauty of this procedure is that one need not take large samples from the underlying distributions F and $G_{(n)}$, compute the components and then apply the subset chi-square test. One need only simulate the limiting distribution of the M components to obtain the asymptotic power function.

The simulation is conducted in the same manner as that which generated Figures 20 and 21. For each set of M iid N(0,1) realizations, an indicator function is set to 1 for rejection and 0 otherwise at each $\gamma = \gamma_i = (i-1)/35$ for $i = 1, \ldots, 36$. These individual functions are then averaged over 10,000 realizations to arrive at the estimated power function. A confidence interval for any point along the estimated function having at least a 95% coverage probability has a half-width of

$$z_{0.975} \cdot \sqrt{\frac{1}{4 \cdot 10{,}000}} \approx 0.0098.$$

There are numerical methods of approximating the power functions of φ_h^2, CVM, and AD so one needn't resort to simulation methods. Each of these statistics is representable as a weighted infinite sum of squares of independent normal random variables under H_o and local alternatives. Under H_o these normal random variables are iid N(0,1); under local alternatives they have nonzero means. Since the weights on the squared normals decrease very rapidly, these numerical methods truncate the infinite series at some point Q. The distribution of

$$T = \sum_{j=1}^{\infty} \theta_j Z_j^2$$

is approximated by

$$T_Q = \sum_{j=1}^{Q} \theta_j Z_j^2.$$

Reflect back to the discussion of the space spanned by the eigenfunctions in Section 3. The distribution of φ_h^2 is to be approximated by the distribution of the projection of the process $\mathrm{KDPo}_{N,h}$ onto a subspace spanned by the eigenfunctions. For the purposes of the approximation, it is clear that whether or not the eigenfunctions form a complete basis for S_D^h is irrelevant.

Durbin and Knott (1972) take the approach described here in finding the distribution of various elements of the CVM statistic. They do add one more term, aX, where X is distributed as χ_ν^2 and is independent of $Z_1, \ldots, Z_Q$. They choose a and ν so that T and T_Q have the same mean and variance.

The approach taken by Durbin and Knott is adopted here. They invert the characteristic function of T_Q by numerical methods. The methods used by Durbin and Knott were originally proposed by Imhof (1961) and Slepian (1958). These methods are geared specifically to quadratic forms of normal random variables. They return the distribution function of T_Q, $F(x)$, given the θ_j's. They are somewhat tedious in that one must perform numerical integration for each x for which $F(x)$ is desired. If the entire distribution function is needed, this can result in quite a lot of computation. A different method is used here; one that applies to a much broader range of cases than the methods of Imhof and Slepian. This new method returns the density function at a range of values, not just at a single value.

This method uses the fast Fourier transform (FFT) to numerically invert the characteristic function. As obvious as this idea sounds, it doesn't appear in the literature in this form. Silverman (1982) [see also Jones and Lotwick (1984)] describes an algorithm which uses the FFT to numerically invert the characteristic function of a kernel density estimate. Otherwise, the FFT has not been used in this manner.

To describe the algorithm, start by assuming $f(x)$ is a continuous density function and that $\phi_X(t)$ is its characteristic function. Then one has [see Parzen (1962b), page 12]

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi_X(t) dt$$

$$\approx \frac{1}{2\pi}\int_{-M}^{M} e^{-itx}\phi_X(t)dt$$

$$\approx \frac{1}{2\pi}\frac{2M}{N}\sum_{j=0}^{N-1} e^{-ix[-M+2Mj/N]}\phi_X(-M+2Mj/N)$$

$$= e^{iMx}\frac{1}{\pi}\frac{M}{N}\sum_{j=0}^{N-1} e^{-2xMji/N}Z(j), \tag{4.2.5}$$

where $Z(j) = \phi_X(-M+2Mj/N)$. The integers M and N are unrelated to their earlier uses. There are two sources of error here, that due to truncation and that due to approximation.

Consider the relation of (4.2.5) to the inverse FFT of $Z(0),\ldots,Z(N-1)$. The inverse FFT is

$$I(k) = \sum_{j=0}^{N-1} e^{-2\pi ijk/N}Z(j), \tag{4.2.6}$$

for $k = 0,\ldots,[N/2]$. Comparing (4.2.5) and (4.2.6), one sees that they are almost the same. Equating the exponents of the two exponentials,

$$\frac{2xMj}{N} = \frac{2\pi jk}{N},$$

$$x = \frac{\pi k}{M},$$

for $k = 0,\ldots,[N/2]$. This defines the x values at which an approximation results. Substituting these x values into e^{iMx}, one finds that

$$e^{iMx} = e^{iM\pi k/M}$$

$$= e^{i\pi k} = \begin{cases} 1 & k \text{ even}, \\ -1 & k \text{ odd}. \end{cases}$$

The estimate $\hat{f}_{NM}$ of $f(x)$ at the points $x_k = \pi k/M$, $k = 0,\ldots,[N/2]$ is

$$\hat{f}_{NM}(x_k) = \frac{M}{N\pi}\Big|\mathrm{Re}[I(k)]\Big|.$$

The estimate of the distribution function F, call it $\hat{F}_{NM}$, can then be found from $\hat{f}_{NM}$ by numerical integration. Using the trapezoidal rule, the estimate works out to be

$$\hat{F}_{NM}(x_k) = \frac{\pi}{M}\left[\sum_{j=1}^{k} f(x_j) - \frac{1}{2}f(x_1) - \frac{1}{2}f(x_k)\right]. \tag{4.2.7}$$

In this work, it is known that the distributions have support on the positive reals so that getting $f(x)$ at $x = \pi k/M$ is not a difficulty. In other work, one could use the forward FFT to get the density at $x = -\pi k/M$. Or one could use the characteristic function of $X \pm b$ to slide the areas of interest within the range of the x_k values.

The approximation used in (4.2.5) appears to be the composite rectangular rule, but it is not. Because of the symmetries of the characteristic function, (4.2.5) is equivalent to the trapezoidal rule. Let

$$\begin{aligned} g(t,x) &= \mathrm{Re}\left[e^{-itx}\phi_X(t)\right] \\ &= \cos tx \cdot \mathrm{Re}[\phi_X(t)] + \sin tx \cdot \mathrm{Im}[\phi_X(t)], \end{aligned}$$

so that $f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} g(t,x)dt$. Note that $g(-t,x) = g(t,x)$ for all x and t. The trapezoidal rule for integrating $g(t,x)$ with respect to t is

$$\begin{aligned} \hat{f}_{NM}(x) &= \frac{M}{N\pi}\left[\sum_{j=0}^{N} g(-M + 2Mj/N, x) - \frac{1}{2}g(-M,x) - \frac{1}{2}g(M,x)\right] \\ &= \frac{M}{N\pi}\sum_{j=0}^{N-1} g(-M + 2Mj/N, x), \end{aligned}$$

since $g(-M,x) = g(M,x)$. This last sum is precisely (4.2.5).

The errors in equations (4.2.5) and (4.2.7) would seem to be working against each other. To make (4.2.5) more accurate one wants M/N small; to make (4.2.7) more accurate one wants M large which means N may need to be huge to make M/N small. One must also select M large enough so that truncation errors in the original integral defining f don't dominate. These forces can be balanced, however. Theorem 4.2.2 gives a numerical consistency result for the approximation $\hat{F}_{NM}$.

Theorem 4.2.2. *Let $\phi_X(t)$ be the characteristic function of f with support on the positive reals and $g(t,x)$ be as defined above. Suppose that $g(t,x)$ is twice differentiable in t (except possibly at t=0, in which case the left and right derivatives must exist) and that*

$$\sup_{t\neq 0, x\in[0,b]} \left|\frac{\partial}{\partial t^2} g(t,x)\right| < \infty$$

and

$$\int_{-\infty}^{\infty} |t^2 \mathrm{Re}[\phi_X(t)]| dt < \infty,$$
$$\int_{-\infty}^{\infty} |t^2 \mathrm{Im}[\phi_X(t)]| dt < \infty.$$

Suppose also that $M^3/N^2 \to 0$ *as* $M, N \to \infty$, *then*

$$\hat{F}_{NM}(b) \to F(b)$$

as $N, M \to \infty$ *for* $b > 0$.

Theorem 4.2.2 says not only that the estimate of the distribution function converges but also gives a bound on the relation of M to N. Rates of convergence are harder to derive. They are probably less useful too since such rates are upper bounds and usually not very good ones. In applying this procedure, one typically observes that if $\hat{f}_{NM}(x_k) < f(x_k)$ then $\hat{f}_{NM}(x_{k+1}) > f(x_{k+1})$. The trapezoidal rule applied to $\hat{f}_{NM}$ gives a result remarkably close to that which would result if f had been used. The procedure seems quite robust to the choice of N and M. However, if really bad choices are made the result is usually quite apparent. The estimated density $\hat{f}_{NM}$ is very wild looking and doesn't come close to integrating to 1.

To test of this procedure, consider using it to invert the characteristic function of the CVM statistic under H_o. In Lemma 4.2.2, the characteristic function $\phi_Q(t)$ of $T_Q = \sum_{j=1}^{Q} \theta_j Z_j^2$ is given.

Lemma 4.2.2. *Let* $Z_1, \ldots, Z_Q$ *be independent with* $Z_j \sim \mathrm{N}(b_j, 1)$. *Let* $\theta_1, \ldots, \theta_Q$ *be a sequence of constants. Then the characteristic function of*

$$\sum_{j=1}^{Q} \theta_j Z_j^2$$

is

$$\phi_Q(t) = \prod_{j=1}^{Q} \left(\frac{1}{1 - 2it\theta_j} \right)^{1/2} \exp[c_j^2 it/(1 - 2it\theta_j)], \tag{4.2.8}$$

where $c_j = \sqrt{\theta_j} b_j$.

Equation (4.2.8) is not in a useful form. Through purely formal manipulations, one arrives at

$$\frac{1}{1-2it\theta_j} = \frac{1}{1+4\theta_j^2 t^2} + i\frac{2\theta_j t}{1+4\theta_j^2 t^2}.$$

One also needs the square root of $a+bi$ which is $c+di$, where

$$d = \sqrt{\frac{1}{2}(-a+\sqrt{a^2+b^2})}$$

$$c = \sqrt{\frac{1}{2}(a+\sqrt{a^2+b^2})}.$$

At this point, equations (4.2.6) and (4.2.7) can be implemented on a computer.

The procedure is run twice, once with a truncation point of $Q = 20$ and once with $Q = 32$. In each case, one more term is added so that the mean of the truncated sum is the same as that of the infinite sum. The results are given in Table 7 and are compared to Anderson and Darling's (1952) values. The values are compared in the quantile domain, not the distribution domain. They are compared in this domain since people actually using the test will want a critical value from the quantile. The large maximum percentage difference observed in Table 7 occurs near the lower endpoint of the distribution where the values of the quantiles are near zero. Larger percentage errors can be forgiven here. The absolute error is small throughout and the maximum percentage error for $u \geq 0.25$ is extremely good.

The procedure has been found to work less well on densities with singular points or large discontinuities. That it should work less well with densities with singularities is not surprising. Most of the simple numerical integration techniques will fail in such cases. The problem with discontinuities comes in inverting the characteristic function. If $f(x)$ is discontinuous at $x = a$, the inversion routine wants to return a value of $[f(a^-) + f(a^+)]/2$ at $x = a$. If the discontinuity is large, this tends to cause the next integration routine to underestimate F. If the point of discontinuity is known and $f(a^-) = 0$, one can always double $\hat{f}_{NM}(a)$.

Table 7

FFT approximation to the quantile function of the Cramér-von Mises statistic under H_o *compared to Anderson and Darling's* (1952) *values. Unless otherwise stated maxima are taken at a grid of* u*'s between* 0.01 *and* 0.999.

	No. of Terms (Q)	
	20	32
$\max\|Q_{NM}(u) - Q(u)\|$	0.0029	0.0019
$\max\|Q_{NM}(u) - Q(u)\|/Q(u)$	11.7%	7.6%
$\max\|Q_{NM}(u) - Q(u)\|$, $u \geq 0.25$	0.0011	0.00025
$\max\|Q_{NM}(u) - Q(u)\|/Q(u)$, $u \geq 0.25$	0.18%	0.10%
M	400	400
N	2048	2048
$Q(0.01)$	0.0248	0.0248
$Q(0.999)$	1.1679	1.1679

The FFT method seems a very good contender to existing techniques both in terms of accuracy and speed. This method is also applicable to a far greater number of cases than the methods of Imhof and Slepian.

4.2.4. *Checking the Conditions of Theorem 4.2.1.* This subsection looks at the details that are involved in showing the conditions of Theorem 4.2.1 are met. These are not shown for all four distributions that are being worked with. The steps are very similar for each and somewhat tedious. Since the Cauchy distribution is widely used as the exception to statistical rules, these conditions are shown for Cauchy location and scale alternatives.

First one needs to find a sequence of constants, $\{a_{(N)}\}$ such that

$$P[-a_{(N)} \leq X_1, \ldots, X_m, Y_{n1}, \ldots, Y_{nn} \leq a_{(N)}] \to 1$$

as $m \wedge n \to \infty$. This is equivalent to the condition bounding the sample quantile function. Since $X_1, \ldots, X_m$ are iid F and $Y_{n1}, \ldots, Y_{nn}$ are iid $F(x - \gamma/\sqrt{n})$ and the two samples are independent, it follows that:

$$(4.2.9) \quad P[-a_{(N)} \leq X_1, \ldots, X_m, Y_{n1}, \ldots, Y_{n1} \leq a_{(N)}]$$
$$= [F(a_{(N)}) - F(-a_{(N)})]^m$$

$$\cdot [F(a_{(N)} - \gamma/\sqrt{n}) - F(-a_{(N)} - \gamma/\sqrt{n})]^n$$
$$= \left[\frac{1}{\pi}\tan^{-1} a_{(N)} - \frac{1}{\pi}\tan^{-1}(-a_{(N)})\right]^m$$
$$\cdot \left[\frac{1}{\pi}\tan^{-1}(a_{(N)} - \gamma/\sqrt{n}) - \frac{1}{\pi}\tan^{-1}(-a_{(N)} - \gamma/\sqrt{n})\right]^n$$
$$= \left[\frac{1}{\pi}\tan^{-1}(a_{(N)} - \gamma/\sqrt{n}) - \frac{1}{\pi}\tan^{-1}(a_{(N)} + \gamma/\sqrt{n})\right]^n$$
$$\cdot \left[\frac{2}{\pi}\tan^{-1} a_{(N)}\right]^m,$$

since $F(x) = 1/2 + (1/\pi)\tan^{-1} x$. Abramowitz and Stegun (1964), page 81, give the following series representation for $\tan^{-1}$:

$$\tan^{-1} x = \frac{\pi}{2} - \frac{1}{x} + \sum_{j=1}^{\infty} \frac{(-1)^{j+1}}{(2j+1)x^{2j+1}}, \quad x \geq 1. \tag{4.2.10}$$

Thus

$$\mathrm{P}[-a_{(N)} \leq X_1, \ldots, X_m, Y_{n1}, \ldots, Y_{n1} \leq a_{(N)}]$$
$$= \left[1 - \frac{2}{\pi a_{(N)}} + o(a_{(N)})\right]^m \left[1 - \frac{1}{\pi(a_{(N)} - \gamma/\sqrt{n})} - \frac{1}{\pi(a_{(N)} + \gamma/\sqrt{n})}\right]^n$$
$$= \left[1 - \frac{2}{\pi a_{(N)}} + o(a_{(N)})\right]^N,$$

since

$$\frac{1}{a_{(N)} - \gamma/\sqrt{n}} + \frac{1}{a_{(N)} + \gamma/\sqrt{n}} = \frac{2}{a_{(N)}} + o(a_{(N)}).$$

Using the fact that $x/(x+1) \leq \ln(1+x) \leq 1/x$ for $x > 0$ one can show that if $a_{(N)} = N^{1+\epsilon}$ then

$$\left[1 - \frac{2}{\pi a_{(N)}} + o(a_{(N)})\right]^N \to 1,$$

as $N \to \infty$. Let $a_{(N)} = N^2$. Next it must be shown that $Q^{H_{(N)}}(1 - 1/N) \leq N^2$. Of course, $Q^{H_{(N)}}(1 - 1/N) \leq N^2$ if and only if $H_{(N)}(N^2) \geq 1 - 1/N$. Since $F(x - \gamma/\sqrt{n}) < F(x)$, it is sufficient to check that $F(N^2 - \gamma/\sqrt{n}) \geq 1 - 1/N$. It can be shown by induction that each partial sum in the series in (4.2.10) is nonnegative which means that

$$\tan^{-1} x \geq \frac{\pi}{2} - \frac{1}{x}, \quad x \geq 1,$$

and thus

$$\begin{aligned} F(N^2 - \gamma/\sqrt{n}) &\geq \frac{1}{2} + \frac{2}{\pi}\Big(\frac{\pi}{2} - \frac{1}{N^2 - \gamma/\sqrt{n}}\Big) \\ &= 1 - \frac{2}{\pi(N^2 - \gamma/\sqrt{n})} \\ &\geq 1 - \frac{1}{N}, \end{aligned}$$

for N sufficiently large. Next the conditions on the sup and inf of the likelihood ratio on the range $[-a_{(N)}, a_{(N)}]$ must be checked. The likelihood ratio is

$$\frac{f(x)}{f(x - \gamma/\sqrt{n})} = \frac{1 + (x - \gamma/\sqrt{n})^2}{1 + x^2},$$

which has extrema at

$$x^* = \frac{1}{2}(\gamma/\sqrt{n} \pm \sqrt{\gamma^2/n + 4}).$$

It is quite clear that $f(x^*)/f(x^* - \gamma/\sqrt{n}) \to 1$ as $m \wedge n \to \infty$. The endpoints of the range should be examined too, since calculus-based methods might miss the endpoints misbehaving:

$$\frac{f(N^2)}{F(N^2 - \gamma/\sqrt{n})} = \frac{1 + (N^2 - \gamma/\sqrt{n})^2}{1 + N^4} \to 1,$$

as $m \wedge n \to \infty$. Therefore, all the conditions are met for the Cauchy location alternative.

Showing that the sample quantile is bounded for local scale alternatives proceeds in a perfectly analogous fashion. Likewise, showing that $F(N^2/|1 + \gamma/\sqrt{n}|) \geq 1 - 1/N$ is carried out in the same fashion. This leaves checking the sup and inf of the likelihood ratio. The likelihood ratio for the scale alternative is

$$\frac{f(x)}{f(x/|1 + \gamma/\sqrt{n}|)} = \frac{1 + (x/|1 + \gamma/\sqrt{n}|)^2}{1 + x^2}.$$

This ratio has an extremum at $x = 0$ and $f(0)/f(0) = 1$. Again, the tails must also be checked:

$$\frac{f(N^2)}{f(N^2/|1 + \gamma/\sqrt{n}|)} = \frac{1 + N^4/(1 + \gamma/\sqrt{n})^2}{1 + N^4} \to 1,$$

as $m \wedge n \to \infty$. Therefore, all the conditions are met for the Cauchy scale alternative.

It has been shown that the conditions for the weak convergence of the empirical comparison distribution process are met for Cauchy location and scale alternatives. The demonstrations for other distributions follow in a completely analogous manner. In the next subsection, power curves and ARE's are found.

4.2.5. *Power Curves and Asymptotic Relative Efficiencies.* Power curves for local location and scale alternatives for the normal and Cauchy distributions are given in this subsection. Power curves are also calculated for two Fourier alternatives. The subset chi-square procedure is seen to perform well, particularly as the alternative stresses higher components. An investigation is made on the effect of the choice of bandwidth on the subset chi-square test. A similar investigation is conducted for φ_h^2. The asymptotic relative efficiencies of the first two components to standard rank statistics are found for location and scale alternatives for four underlying distributions: normal, Cauchy, logistic, and Laplace. The efficiency of the components is seen to vary with the bandwidth. For the distributions considered, a larger bandwidth tends to do better for location alternatives and a smaller one better for scale alternatives.

For the subset chi-square test, a cutoff of 0.001 is used for including components in the test. Thus, for $h = 0.5, 0.4, 0.3, 0.2$, and 0.1, a truncation point of $M = 4, 6, 8, 12$, and 23 is used, respectively (cf. Table 6). Table 6 would say that for $h = 0.5$ that $M = 5$ should be used but it was not. The eigenvalue for the fifth component in this case is 0.002 so that its exclusion should not be significant. All power curves are derived for $\lambda_0 = 0.5$ and $0 \leq \gamma \leq 7$.

Figures 26 and 27 present power curves for normal and Cauchy location alternatives, respectively. For the normal case, the techniques arrange themselves as follows from highest to lowest power: AD, CVM, $\varphi_{0.5}^2$, subset chi-square: h =0.5, 0.3, 0.2. There is a gap between the top and bottom three. This is not unexpected considering previous remarks on the behavior of these statistics. The components are down-weighted at such a rate that the first few dominate. The normal location alternative affects mainly the first component.

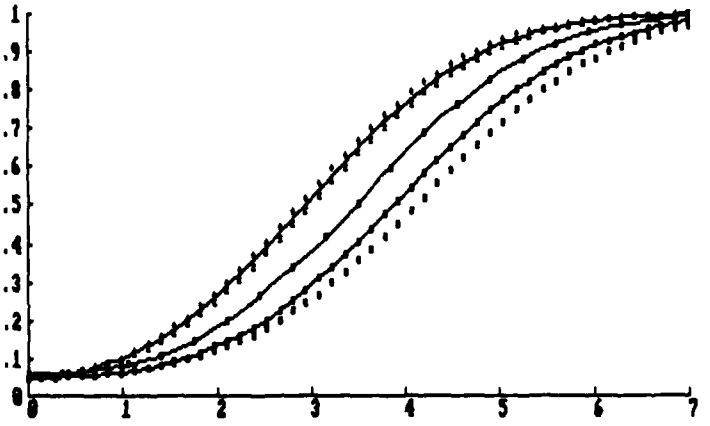

Fig. 26. *Power of the subset chi-square,* $\varphi^2_{0.5}$, CVM, *and* AD *tests against normal location alternatives. The solid line is CVM; the* + *'s are AD; the* x *'s are* $\varphi^2_{0.5}$*; the solid line with sparse blocks is the subset chi-square,* $h = 0.5$*; the solid line with dense blocks is the subset chi-square,* $h = 0.3$*; the blocks are the subset chi-square,* $h = 0.2$.

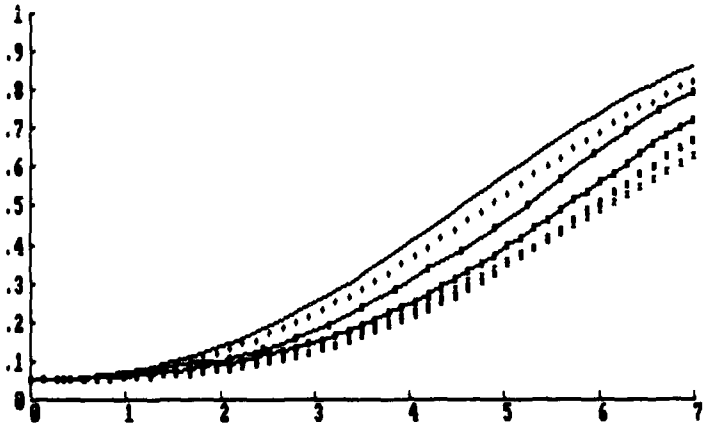

Fig. 27. *Power of the subset chi-square,* $\varphi^2_{0.5}$, CVM, *and* AD *tests against Cauchy location alternatives. The solid line is CVM; the* + *'s are AD; the* x *'s are* $\varphi^2_{0.5}$*; the solid line with sparse blocks is the subset chi-square,* $h = 0.5$*; the solid line with dense blocks is the subset chi-square,* $h = 0.3$*; the blocks are the subset chi-square,* $h = 0.2$.

Table 8 presents the efficacies of the components for normal and Cauchy location and scale alternatives. One can see that for the normal location alternative, the first component has the largest efficacy.

For the Cauchy location alternative, the power functions are much closer. In fact, $\varphi^h_{0.5}$ does worse than the subset chi-squares. Referring to Table 8 again, one sees that the Cauchy location alternative places the most weight on the third component and about half as much on the first. This is unusual for a location alternative, yet recall from Figure 1(c) that the comparison density for this case is not monotone. For a bandwidth of $h = 0.5$, the third component is downweighted severely in $\varphi^2_{0.5}$ (cf. Figure 23), hence its poor performance. It will also been seen later in the subsection that the components making up CVM are more efficient against this alternative than those making up $\varphi^2_{0.5}$.

The situation changes even further for scale alternatives. Figures 28 and 29 give the power functions for normal and Cauchy scale alternatives, respectively. Table 8 verifies that these alternatives principally affect the second component. The first component has no influence at all and so the statistics CVM, AD, and φ^2_h drop off. In both these cases the subset chi-square ($h = 0.5$) is most powerful. The $\varphi^2_{0.5}$ statistic is next followed by the two subset chi-squares ($h = 0.3, 0.1$). The $\varphi^2_{0.5}$ statistic shows relative improvement from the Cauchy location alternative for two reasons. First, the second component receives more weight than the third component which dominated the Cauchy location alternative. Second, as shall be seen, the second component performs much better against both these alternatives than the first does against Cauchy location alternatives.

Figure 30 presents the power curves for what shall be called Fourier alternative 1. Referring to equation (4.2.4), Fourier alternative 1 is defined by $k = 5$ and $a' = 2.5 \cdot (-.4, -.5, .6, 1, 1)/30$. These coefficients are chosen so that the Wilcoxon, median, and Mood tests all have power equal to their size. That is, these tests are no better than one which randomly rejects H_o 100α percent of the time. From Table 9 it can be seen that this alternative affects mainly the third, fourth, and fifth components. For all but the largest bandwidths, the first component is involved as well. The significance of the weights used by the

Table 8

Efficacies of the components of the kernel density process, normal and Cauchy location and scale alternatives, $\lambda_0 = 0.5$ *and* $\gamma = 1$.

Component	Bandwidth 0.1	0.2	0.3	0.4	0.5
			Normal Location		
1	0.621	0.672	0.685	0.691	0.692
2	0.000	0.000	0.000	0.000	0.000
3	0.276	0.180	0.141	0.102	0.058
4	0.000	0.000	0.000	0.000	0.000
5	0.138	0.094	0.022	0.053	0.094
6	0.000	0.000	0.000	0.000	0.000
7	0.094	0.023	0.050	0.072	0.061
8	0.000	0.000	0.000	0.001	0.000
			Normal Scale		
1	0.000	0.000	0.000	0.000	0.000
2	0.967	0.932	0.899	0.871	0.848
3	0.000	0.000	0.000	0.000	0.000
4	0.015	0.093	0.143	0.199	0.266
5	0.000	0.000	0.000	0.000	0.000
6	0.016	0.063	0.173	0.240	0.337
7	0.000	0.000	0.000	0.000	0.000
8	0.018	0.109	0.151	0.198	0.142
			Cauchy Location		
1	0.149	0.220	0.249	0.273	0.300
2	0.000	0.000	0.000	0.000	0.000
3	0.475	0.447	0.433	0.418	0.399
4	0.000	0.000	0.000	0.000	0.000
5	0.040	0.035	0.007	0.021	0.010
6	0.000	0.000	0.000	0.000	0.000
7	0.017	0.003	0.006	0.011	0.002
8	0.000	0.000	0.000	0.000	0.000
			Cauchy Scale		
1	0.000	0.000	0.000	0.000	0.000
2	0.374	0.426	0.448	0.464	0.480
3	0.000	0.000	0.000	0.000	0.000
4	0.302	0.249	0.220	0.185	0.139
5	0.000	0.000	0.000	0.000	0.000
6	0.108	0.075	0.017	0.014	0.000
7	0.000	0.000	0.000	0.000	0.000
8	0.064	0.013	0.011	0.000	0.002

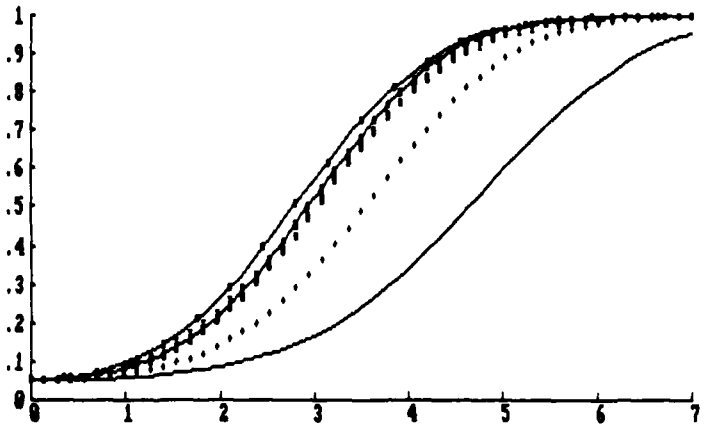

Fig. 28. *Power of the subset chi-square,* $\varphi^2_{0.5}$, CVM, *and* AD *tests against normal scale alternatives. The solid line is CVM; the* $+$*'s are AD; the* x*'s are* $\varphi^2_{0.5}$*; the solid line with sparse blocks is the subset chi-square,* $h = 0.5$*; the solid line with dense blocks is the subset chi-square,* $h = 0.3$*; the blocks are the subset chi-square,* $h = 0.2$.

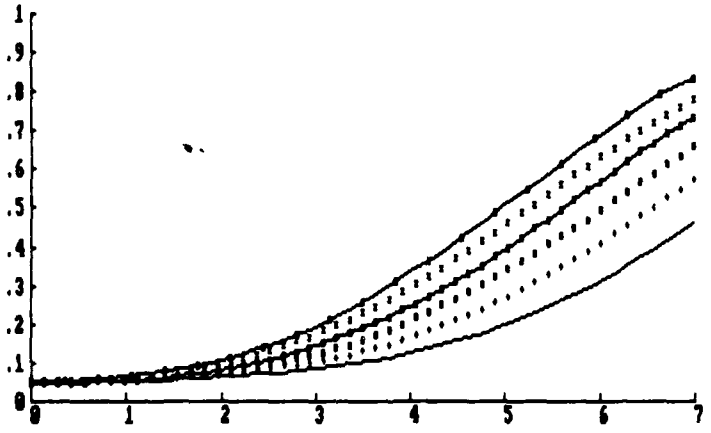

Fig. 29. *Power of the subset chi-square,* $\varphi^2_{0.5}$, CVM, *and* AD *tests against Cauchy scale alternatives. The solid line is CVM; the* $+$*'s are AD; the* x*'s are* $\varphi^2_{0.5}$*; the solid line with sparse blocks is the subset chi-square,* $h = 0.5$*; the solid line with dense blocks is the subset chi-square,* $h = 0.3$*; the blocks are the subset chi-square,* $h = 0.2$.

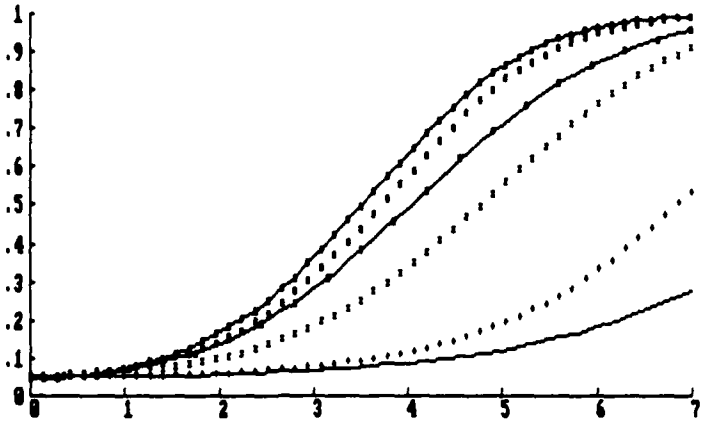

Fig. 30. *Power of the subset chi-square,* $\varphi^2_{0.3}$, CVM, *and* AD *tests against Fourier alternative 1. The solid line is CVM; the* $+$*'s are AD; the* x*'s are* $\varphi^2_{0.3}$*; the solid line with sparse blocks is the subset chi-square,* $h = 0.5$*; the solid line with dense blocks is the subset chi-square,* $h = 0.3$*; the blocks are the subset chi-square,* $h = 0.2$.

Table 9

Efficacies of the components of the kernel density process, Fourier alternatives, $\lambda_0 = 0.5$ *and* $\gamma = 1$.

	Bandwidth				
Component	0.1	0.2	0.3	0.4	0.5
	Fourier Alternative 1				
1	0.272	0.245	0.216	0.138	0.062
2	0.118	0.090	0.087	0.068	0.016
3	0.320	0.360	0.395	0.412	0.366
4	0.467	0.484	0.489	0.510	0.526
5	0.425	0.441	0.452	0.514	0.564
6	0.000	0.000	0.000	0.000	0.000
7	0.000	0.000	0.000	0.000	0.000
8	0.000	0.000	0.000	0.000	0.000
	Fourier Alternative 2				
1	0.047	0.039	0.053	0.099	0.073
2	0.008	0.019	0.026	0.087	0.063
3	0.165	0.160	0.180	0.235	0.228
4	0.317	0.308	0.278	0.343	0.458
5	0.225	0.263	0.319	0.165	0.004
6	0.746	0.731	0.737	0.746	0.658
7	0.681	0.628	0.667	0.682	0.715
8	0.000	0.000	0.000	0.000	0.000

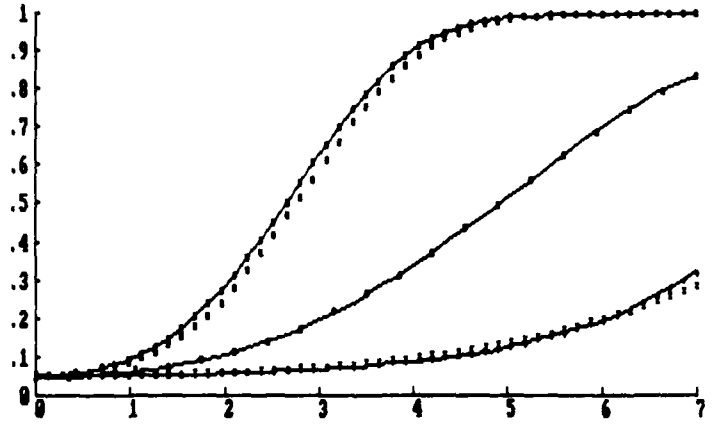

Fig. 31. *Power of the subset chi-square,* $\varphi^2_{0.3}$, CVM, *and* AD *tests against Fourier alternative 2. The solid line is CVM; the* $+$*'s are AD; the* x*'s are* $\varphi^2_{0.3}$*; the solid line with sparse blocks is the subset chi-square,* $h = 0.5$*; the solid line with dense blocks is the subset chi-square,* $h = 0.3$*; the blocks are the subset chi-square,* $h = 0.2$.

statistics φ^2_h, CVM, and AD is beginning to become clear. The subset chi-square tests for each bandwidth do substantially better than the traditional statistics CVM and AD. The $\varphi^2_{0.3}$ statistic does improve on these two considerably, but its first component does have a fair sized efficacy for this alternative. Note that the power of the subset chi-square test no longer decreases with the bandwidth. The ordering is h =0.3, 0.1, 0.5.

Figure 31 presents what shall be called Fourier alternative 2. Again, in reference to equation (4.2.4), this alternative is defined by $k = 7$ and $a' = (.1978, .3208, -.9395, -1.308, -.1373, 1, 1)/15$. These coefficients are chosen so that the Wilcoxon, median, normal scores (location), Mood, and normal scores (scale) tests all have power equal to their size. From Table 9, it can be seen that this alternative affects mainly the sixth and seventh components. This case is even more extreme than the last. The subset chi-square with $h = 0.3$ and 0.1 do very well. The CVM, AD, and φ^2_h statistics perform uniformly poorly. The subset chi-square with $h = 0.5$ is between these two sets. The ordering of the power of subset chi-square test by bandwidth is the same as for Figure 30.

At this point a word about the effect of the bandwidth on the power of the subset chi-square test is in order. For scale and location alternatives, the order is according to decreasing bandwidth. As the alternatives move to higher components, this order changes. Including more components that don't have much (any) efficacy reduces the power of the test. For location and scale alternatives, only the first two or three components are important. Reducing the bandwidth adds components to the decision process which carry little or no signal (efficacy). This translates to a reduction in power. As the alternative moves to higher components, the larger bandwidth excludes components that carry the signal. That is, the larger bandwidths simply don't consider alternatives in these directions. The larger bandwidths then start to be less powerful than the smaller bandwidths.

These observations strike at the heart of the choice of truncation point. If one chooses M too large (h too small) then power decreases because one is adding noise to the process. If one chooses M too small (h too large) the test also loses power because components with significant efficacy are not considered. In the worst case the test would be inconsistent if the alternative did not affect the first M components at all. It is believed that by choosing the bandwidth carefully in the initial stage these extremes can be avoided. This procedure is certainly preferable to the alternative of using a standard statistic. In that case, one is assured of poor performance for alternatives stressing higher components.

The effect of the bandwidth on φ_h^2 is less clear-cut. Figures 32 and 33 present power curves for φ_h^2 with h =0.5, 0.3, and 0.1 for normal location and scale alternatives, respectively. The ordering here is more complex. For the location alternative, the order from most to least powerful is h =0.5, 0.3, and 0.1; for the scale alternative, it is h =0.3, 0.1, 0.5. The location alternative is easier to explain: $h = 0.5$ is the most efficient first component (as will be seen) and it gives the least weight to other components. For scale alternatives, although $h = 0.1$ is the most efficient bandwidth, $h = 0.3$ is not much worse (again, as shall be seen). However, $h = 0.1$ gives much greater weight to many more components (recall Figure 14). This added variability causes $\varphi_{0.1}^2$ to be

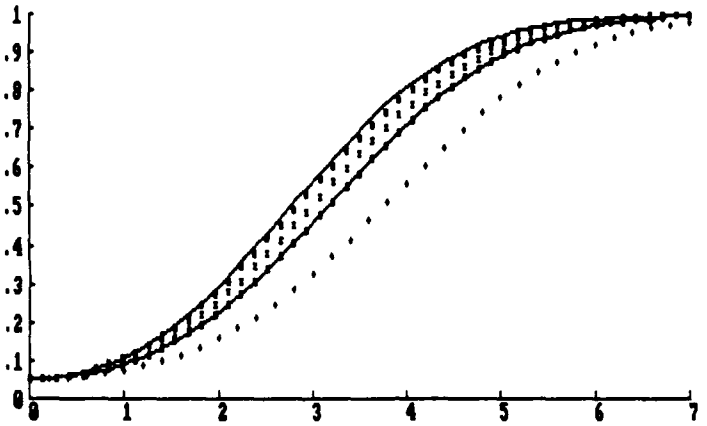

Fig. 32. *Power of* φ_h^2, *t-test, and first component* $(h = 0.5)$ *against normal location shifts. The* x*'s are* $\varphi_{0.5}^2$*; the solid line with dense blocks is* $\varphi_{0.3}^2$*; the* $+$*'s are* $\varphi_{0.1}^2$*; the solid line is the t-test; the blocks are the first component* $(h = 0.5)$.

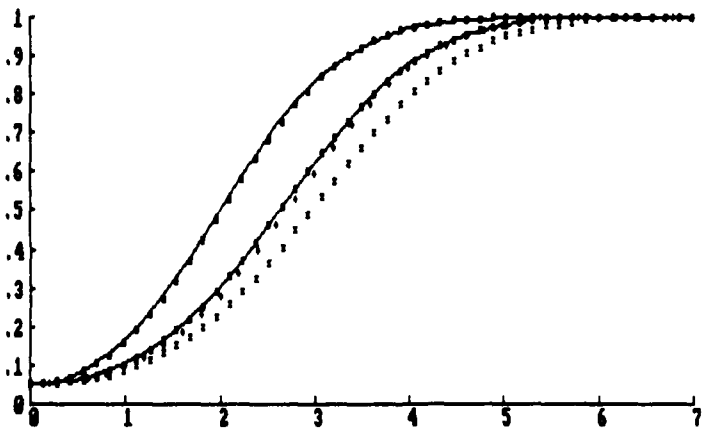

Fig. 33. *Power of* φ_h^2, *F-test, and second component* $(h = 0.1)$ *against normal scale shifts. The* x*'s are* $\varphi_{0.5}^2$*; the solid line with dense blocks is* $\varphi_{0.3}^2$*; the* $+$*'s are* $\varphi_{0.1}^2$*; the solid line is the t-test; the blocks are the second component* $(h = 0.1)$.

less powerful than $\varphi^2_{0.3}$.

Figures 32 and 33 each include two more power curves. Figure 32 also gives the power curve for the t-test and the first component ($h = 0.5$). Figure 33 includes the power curves for the F-test and the second component ($h = 0.1$). These curves illustrate several statements made in Section 2. The first is that a test against a more specific alternative hypothesis will tend to be more powerful. The tests of the first and second components do better than any of the portmanteau tests for testing location and scale shifts, respectively. Of course, if one used the first component to test a scale alternative one would find it did miserably. This fact is clearly demonstrated in Table 8: the first component has efficacy equal to 0 for both the scale alternatives. The second point is that asymptotically, nonparametric tests can do just as well as parametric tests. The first and second components are not the optimal scores for shifts in the normal distribution (the normal scores are). Yet they do very well, indeed.

Tables 10 and 11 give the asymptotic relative efficiencies of the first two components to standard rank statistics. Table 10 gives ARE's of the first component to the Wilcoxon, median, normal scores (location), and cosine tests. These are all tests for location. Table 11 gives the ARE's of the second component to Mood, normal scores (scale), and cosine tests. These are all tests for scale. The component is more efficient than the standard rank statistic if the entry exceeds 1. The score functions for all the rank tests but the cosine are given in Table 2. The score functions for the cosine rank tests are below their column title in Tables 10 and 11. The purpose of including the cosine rank statistics will become clear shortly.

From Table 10 it appears that the standard rank statistics are more efficient than the components for the Cauchy and Laplace distributions. For the normal and logistic distributions, the best component is on a par with the best standard test.

It is apparent from Table 10 that the bandwidth does influence materially the properties of the components. Recalling Figures 15 through 17, the score functions for the first component do materially change with the bandwidth. It is

Table 10

Asymptotic relative efficiencies of the components to standard rank statistics for location alternatives.

	Standard Rank Statistic			
Bandwidth	Wilcoxon	Median	Normal Scores	Cosine $\cos \pi u$
		Normal		
0.5	1.016	1.518	0.970	1.074
0.4	1.011	1.511	0.966	1.069
0.3	0.996	1.489	0.952	1.054
0.2	0.958	1.431	0.915	1.013
0.1	0.817	1.221	0.780	0.864
		Cauchy		
0.5	0.593	0.445	0.827	0.499
0.4	0.491	0.368	0.685	0.413
0.3	0.408	0.306	0.569	0.343
0.2	0.318	0.239	0.444	0.268
0.1	0.146	0.109	0.203	0.123
		Logistic		
0.5	0.939	1.250	0.977	0.950
0.4	0.905	1.204	0.942	0.916
0.3	0.857	1.141	0.892	0.867
0.2	0.781	1.039	0.812	0.790
0.1	0.591	0.786	0.615	0.598
		Laplace		
0.5	0.722	0.543	0.846	0.668
0.4	0.692	0.521	0.811	0.640
0.3	0.653	0.491	0.765	0.604
0.2	0.574	0.432	0.673	0.531
0.1	0.379	0.285	0.444	0.350

Table 11

Asymptotic relative efficiencies of the components to standard rank statistics for scale alternatives.

	Standard Rank Statistic		
Bandwidth	Mood	Normal Scores	Cosine $\cos 2\pi u$
	Normal		
0.5	1.009	0.783	1.345
0.4	1.065	0.826	1.418
0.3	1.134	0.880	1.511
0.2	1.219	0.946	1.625
0.1	1.312	1.019	1.749
	Cauchy		
0.5	1.007	1.611	0.923
0.4	0.940	1.504	0.862
0.3	0.877	1.402	0.804
0.2	0.794	1.270	0.728
0.1	0.613	0.980	0.562
	Logistic		
0.5	1.006	0.902	1.255
0.4	1.042	0.935	1.300
0.3	1.084	0.973	1.353
0.2	1.125	1.010	1.405
0.1	1.132	1.015	1.413
	Laplace		
0.5	1.002	0.900	1.216
0.4	1.029	0.924	1.249
0.3	1.071	0.961	1.300
0.2	1.115	1.001	1.354
0.1	1.129	1.013	1.370

interesting to note that for these four cases a bandwidth of 0.5 is more efficient than 0.1. There are, of course, cases where the ordering should change. Whatever the character of the underlying distribution must be for the order to change, it is not embodied in the examples here. These examples, however, are somewhat restricted. Although they do cover a range of tail behavior, they are all unimodal densities which are symmetric about 0.

The second component is generally more competitive against standard rank statistics than the first as evidenced in Table 11. There is also considerably less variation across bandwidths. This, too, is not surprising if one reflects back to Figures 15 through 17. The character of the score functions of the second component seems to change less than that of the first. A bandwidth of 0.1 is optimal in each case except the Cauchy in which case a bandwidth of 0.5 is preferred.

The change in the best bandwidth across location and scale alternatives for the same distribution is disturbing. This means that one cannot choose the bandwidth to best protect against both location and scale shifts. One can, however, select the bandwidth so that both components do nearly as well as possible for both.

The cosine based rank statistics have been included so that one can compare the method of Section 3 to that of the components of φ^2 advanced by Eubank, LaRiccia, and Rosenstein (1987). Their method is based on using a complete orthonormal basis as a set of score functions; for instance, the cosine basis $\{\cos j\pi u\}$. They do not develop an estimator of the comparison density, nor do they suggest a technique for testing the components. They do discuss the components as testing successively higher frequency departures of the comparison density from uniformity. They also point out that the components are asymptotically iid N(0,1).

Filling in the obvious details not in their paper: if one didn't have a truncation point, M, in mind, one could choose it from the data. The estimate of the comparison density is an ordinary orthogonal series estimate like those discussed in Section 2.

An orthogonal series was not adopted for reasons given in Section 2. One would have to be careful in the choice of basis. Orthogonal series methods can impose constraints on the estimated density. The advantages of components derived from the eigenfunctions were discussed in Section 3.

Admittedly, the methodology based on the boundary kernel is more complex, particularly computationally, but the burden brings with it the convenience of use and unification. It has been said already that the goal is to make the statistician's job difficult so that the researcher's job is easier.

If one desired to compare the two methods from a testing point of view, it would first be necessary to specify the test to be applied to the orthogonal series components. Any method of Subsection 3.3 is applicable. However, there isn't any need to go to all this trouble. The result one would be certain to find is that which is more powerful depends on the alternative being considered. Since both tests are based on the components, the components with the greatest efficacies (and hence ARE's) should yield the more powerful test. Examining the ARE's of the cosine score functions in Tables 10 and 11, one sees that for some alternatives such as the Cauchy location and scale, the cosine scores do relatively better. For other alternatives, such as normal scale and location ($h \neq 0.1$), and logistic and Laplace scale shifts the cosine basis does relatively worse. Certainly, an overall test would reflect these observations.

In summary, power and asymptotic relative efficiencies were examined in this subsection. Power curves were found for location and scale alternatives for the normal and Cauchy distributions. Power curves were also found for two Fourier alternatives. These curves were derived for the subset chi-square test applied to the components, φ_k^2, CVM, and AD. It was observed that as the alternative affected higher components, these last three statistics performed ever more poorly. It was also observed that the most powerful bandwidth is the largest bandwidth which still has a truncation point which picks up the components most important to the alternative under consideration.

The ARE's of the first two components to standard rank tests were found for location and scale alternatives of four underlying distributions: the normal,

Cauchy, logistic and Laplace. The components generally performed better for scale alternatives than location alternatives. The bandwidth was seen to affect the performance of the components. It was also seen one cannot always choose the bandwidth to protect best against both location and scale alternatives for a given underlying distribution.

4.3. Size Studies

4.3.1. *Introduction.* The finite-sample size of the subset chi-square test is investigated in this subsection. Simulations are run to determine the size of the test using both the small sample mean and the asymptotic mean when centering the components. The sizes using the small sample mean are seen to be much better than those found using the asymptotic mean. The sizes for the small sample mean also tend to be below their nominal value. This means that the test is conservative, which is better than being liberal. The estimated sizes using the asymptotic mean tend to be greater than the stated size.

4.3.2. *Size Study.* Simulations are used to estimate the small sample size of the subset chi-square test applied to the components. They are conducted as follows. For each set of sample sizes, m and n, and for each bandwidth and truncation point, R iterations are made. Within each iteration, two independent random samples of sizes m and n are drawn from the U(0,1) distribution. The boundary kernel estimate of the comparison density is found. From it the components are calculated. Either the small sample or asymptotic mean is subtracted, as appropriate. The subset chi-square test is then applied to the components. For the j^{th} iteration, B_j is set to 1 if the test rejects and 0 if not. The size of the test is estimated as

$$\frac{1}{R}\sum_{j=1}^{R} B_j.$$

There is no reuse of the data here as in the simulations to find power curves. These simulations are to find point estimates, not estimates of functions. Reuse of the data serves no purpose in this context. For simulations using the asymptotic

mean, $R = 1000$; for simulations using the small sample mean, $R = 5000$. A greater number of simulations are run using the small sample mean because this is the case of greater interest. The simulations using the asymptotic mean are run to illustrate the improvements that result from using the small sample mean. All subset chi-square tests are conducted at a stated size of 0.05. The simulations are run for $h = 0.5$ $(M = 4)$, $h = 0.3$ $(M = 8)$, and $h = 0.2$ $(M = 12)$. They are run for various combinations of m and n ranging from 5 to 100.

Table 12 gives the results using the small sample means. The sizes look remarkably good. The half width of a joint confidence interval to test $H_o: \alpha_i = 0.05$, for $i = 1, \ldots, 30$ is

$$z_{0.975^{1/30}} \sqrt{\frac{0.05 \cdot 0.95}{5000}} \approx 0.010.$$

The joint test is rejected. Ten of the thirty estimated sizes fall outside the confidence limits. These are $n = m = 5$ for $h = 0.3, 0.2$; $n = 10$, $m = 5$ for $h = 0.3, 0.2$; $n = 100$, $m = 5$ for $h = 0.5, 0.3, 0.2$; $n = m = 10$ for $h = 0.3, 0.2$; and $n = 20$, $m = 10$ for $h = 0.2$. These are the smallest sample sizes and the smallest bandwidths. Only one size corresponding to a bandwidth of 0.5 falls outside the confidence limits.

The sizes which are significantly different from 0.05 fall into one of two categories. The first category is small but nearly equal sample sizes and very small bandwidth. With 5 or 10 observations, nobody would use a bandwidth as small as 0.3 even. In this category are $n = m = 5$; $n = 10$, $m = 5$; $n = m = 10$ and $n = 20$, $m = 10$. Hence, the fact that these sizes are significantly different from the stated size is not that much of an issue. The second category is very unequal sample sizes. This is the case of $n = 100$ and $m = 5$. If this sort of case were to arise, one would need to be aware that the size of the test may be smaller than stated. However, such extreme differences in sample size don't usually arise.

Of the sizes found to be different from 0.05, 8 are below 0.05 and only 2 above. Of all the estimated sizes, 6 are above 0.05 and 24 below. The test seems to err on the side of falling under 0.05. This is good as it means the test is

Table 12

Estimated small sample sizes of the subset chi-square test applied to the components centered by their small sample mean. The stated size of the test is 0.05.

		Number of Components		
m	n	4	8	12
5	5	0.060	0.031	0.117
5	10	0.046	0.035	0.023
5	100	0.036	0.035	0.034
10	10	0.049	0.039	0.024
10	20	0.047	0.042	0.038
10	100	0.042	0.046	0.040
20	20	0.057	0.043	0.040
20	100	0.052	0.049	0.050
50	50	0.048	0.045	0.049
50	100	0.049	0.052	0.047
bandwidth		0.5	0.3	0.2

conservative. When a researcher runs a test at size 0.05, he has decided that he will accept false rejections 5% of the time. It is better that when the test rejects, the probability of a false rejection is less than this figure than above. Taken as a whole, Table 12 presents very encouraging results.

Table 13 presents the estimated sizes of the subset chi-square test applied to the components which are centered by their asymptotic mean. The situation here is much less satisfactory than that above. The half width of a joint confidence interval for testing $H_o: \alpha_i = 0.05$ for $i = 1, \ldots, 33$ is

$$z_{0.975^{1/33}} \sqrt{\frac{0.05 \cdot 0.95}{1000}} \approx 0.022.$$

Again, the global test is rejected. Seven of the sizes fall outside their confidence interval. These are: $m = n = 10$ for $h = 0.3, 0.2$; $m = 20$, $n = 10$ for $h = 0.5, 0.3, 0.2$; and $m = 100$, $n = 10$ for $h = 0.3, 0.2$. Table 13 has fewer significantly different sizes than Table 12 for two reasons. The first is that Table 13 has no sample sizes below 10. Seven of the ten significant sizes in Table 12 have a sample size of 5. The second is that Table 13 is constructed with only 1000 replications and so the estimates are less accurate.

Table 13

Estimated small sample sizes of the subset chi-square test applied to the components centered by their asymptotic mean. The stated size of the test is 0.05.

		Number of Components		
m	n	4	8	12
10	10	0.063	0.076	0.081
10	20	0.051	0.057	0.048
10	100	0.046	0.043	0.039
20	10	0.073	0.084	0.073
20	20	0.059	0.058	0.052
30	30	0.059	0.050	0.057
50	50	0.048	0.051	0.052
50	100	0.052	0.059	0.053
100	10	0.071	0.091	0.085
100	50	0.046	0.052	0.048
100	100	0.049	0.053	0.055
bandwidth		0.5	0.3	0.2

There are two disturbing features about Table 13. First, it is quite clear that the test using the asymptotic mean is not invariant. The discussion in Subsection 3.3.3 predicted that for $m \gg n$, one would reject too often. This is precisely the case observed here. To reach a different conclusion based on which sample is termed the first is extremely undesirable. The second disturbing feature is related to the first. All 7 of the significant sizes are above 0.05; taking the table as a whole, 24 of the 33 (73%) are above 0.05. If the procedure must err, it is preferable for the size to be below what is stated, not above.

There are two conclusions to be drawn from this subsection. The first is that the subset chi-square using the components centered by their small sample means performs very well in terms of keeping its stated size even for small samples. This is true as long as a reasonable bandwidth is used and the samples sizes are not very dissimilar. When it does err, it tends to err on the side of having a smaller than stated size. This is also good. The second is that subtracting the small sample mean instead of the asymptotic mean is a good idea. The test is invariant and the estimated sizes are much closer to the stated value.

In summary of the section, power and size tests have been run. The subset chi-square test applied to the components of the kernel density process was seen to be a credible procedure. It is very competitive with the Cramér-von Mises and Anderson-Darling statistics. These latter two statistics were seen to have very low power against alternatives stressing higher order components. This deterioration starts with the second component. By the time the fourth and higher components make up the alternative, these statistics do very poorly. The subset chi-square test did not exhibit this trait. Instead, the choice of truncation point plays a crucial role in determining its power. If the truncation point is chosen too large, then the test loses power because the signal is lost in the noise. If the truncation point is chosen too small, then the test loses power because the signal is excluded from the test. A careful choice of bandwidth should reduce the likelihood of experiencing these extremes.

5. EXAMPLES AND APPLICATIONS

5.1. Introduction

The two sample procedure derived in Section 3 is applied to two data sets in this section. The first data set is the observed weekly rate of return to two savings and loan institutions over a 103 week period while the second data set is simulated. The two samples are each normally distributed but with different means and variances. This is an example of the well-known Behrens-Fisher problem [see Kendall and Stuart (1979), pages 152 ff.]. The first data set is an example of a case where none of the standard statistics identify a difference in the populations. The second is an example in which all the standard statistics identify a difference in the populations. In the first instance, it will be seen how the subset chi-square test applied to the components can find differences where the others fail. In the second instance, it will be seen how the new methodology can clarify a distinction also found to exist by other methods.

5.2. The Savings and Loan Data

The savings and loan data consist of the weekly rate of return for two New York Stock Exchange listed savings and loans over a 103 week period. The observation period is July 3, 1981 to June 30, 1983. The first sample consists of returns for H.F. Ahmanson and Company; the second consists of returns for Financial Corporation of Santa Barbara. The data are drawn from Standard and Poor's Stock Price Data. The return for week t is defined as

$$R(t) = \ln \frac{P(t)}{P(t-1)},$$

where $P(t)$ is the price in week t. Dividends are added to the price in the week they are paid. This definition is often used in the finance literature; see, for example, Fama (1976) pages 12-20.

The question to be investigated is whether the returns are distributed the same for the two institutions. One might suppose that the distribution of returns

differs by region, solvency, or in reaction to some outside event. Often one is interested in whether returns differ for the same company or industry in two different time periods. These sorts of questions are not infrequently asked in finance. See, for example, Dann and James (1982) or Brown and Warner (1980).

The returns for H.F. Ahmanson and Company are given in Table 14. The returns for Financial Corporation of Santa Barbara are given in Table 15. The data sets are pictured in Figures 34 and 35. Graphed as time series, they have much the same appearance. These data are assumed to be realizations of independent sets of iid random variables. It is important to check the validity of these assumptions. Since these data sets are observed economic time series, it would be reasonable to expect them to be autocorrelated although this effect may well be reduced by taking the differences of the logarithms of the data. Since each savings and loan is subject to similar national economic and regulatory conditions, it is also reasonable to expect the two series not to be independent.

Timeslab [Newton (1988)] is used to determine the correlation structure of each series. In each case, the CAT [Parzen (1977)] criterion chose an autoregressive order of zero. While not a guarantee of independence, the lack of correlation is very good news. The sample cross-correlation coefficient between the two series is calculated as $r = 0.429$. This value is significant at the 1% level.

A positive correlation between the two series would likely reduce the chance of rejecting H_o if it were true since the two series would appear more similar. The extreme case would be if the correlation were 1. Breaking ties by midranking, the ranks would be

$$R_i = \frac{2i - (1/2)}{N},$$

which are extremely uniform. Lower levels of correlation should still lead to more uniform ranks than would otherwise be observed. Hence, tests based on ranks are expected to be conservative in this case.

This analysis is borne out by a simulation study. Two samples of size 100 are drawn from the standard normal distribution. Three levels of cross-correlation are used: 0.447, 0.707, and 0.894. As the savings and loan data appears nearly normally distributed, this choice of distribution and sample sizes should yield

Table 14

Weekly returns for H.F. Ahmanson and Company from July 3, 1981 *to June* 30, 1983. *Table values are multiplied by* 100.

-1.400	0.000	-6.000	1.500	-0.800	-0.800	3.000
-0.700	-0.800	7.300	-2.900	-6.700	-0.800	0.800
0.000	-4.000	-5.000	0.800	5.700	3.900	-1.500
0.000	-7.200	-1.700	1.700	0.800	-4.200	-14.800
-6.200	6.200	-8.300	-11.500	-2.500	16.100	-11.200
1.700	-1.700	0.000	2.400	0.000	2.300	0.000
-5.800	7.000	0.000	-10.700	-6.500	-8.300	2.900
-4.300	1.500	4.300	-2.800	1.400	5.500	-2.700
2.700	3.900	32.500	-3.800	9.200	-2.700	-0.900
3.600	7.600	11.500	1.400	9.500	3.200	26.900
2.800	8.100	2.600	0.000	0.000	0.400	0.400
-7.800	-3.700	16.300	-26.700	9.400	-1.900	6.500
-1.800	16.800	15.100	-11.300	1.500	-6.500	0.800
8.200	7.900	2.000	1.000	0.300	1.900	-4.500
-10.000	1.800	-1.100	-2.200	-7.300		

Table 15

Weekly returns for Financial Corporation of Santa Barbara from July 3, 1981 *to June* 30, 1983. *Table values are multiplied by* 100.

2.740	-2.740	5.407	6.372	-2.500	-6.538	-4.139
-11.955	0.000	6.156	-6.156	-6.560	-5.219	5.219
-5.219	-5.506	-5.827	-4.082	9.909	5.506	0.000
-1.802	-7.551	-4.001	2.020	-2.020	-6.318	-27.329
2.817	-8.701	-3.077	-6.454	-10.536	-3.774	-3.922
0.000	0.000	-8.338	19.671	0.000	-3.637	-25.131
4.652	12.783	7.686	-3.774	-16.705	-4.652	4.652
-4.652	9.097	0.000	4.256	8.004	3.774	-3.774
0.000	3.774	31.508	5.264	-10.821	-8.961	6.063
-2.985	5.884	20.585	6.744	-2.198	2.198	19.671
1.770	-3.572	-9.531	-4.082	6.063	-1.980	-4.082
-15.763	23.767	-10.110	0.000	2.105	0.000	2.062
-2.062	2.062	15.123	-5.407	3.637	17.934	8.577
1.361	7.796	-11.935	0.000	5.481	6.454	-5.129
-4.027	-4.196	-4.380	2.941	4.256		

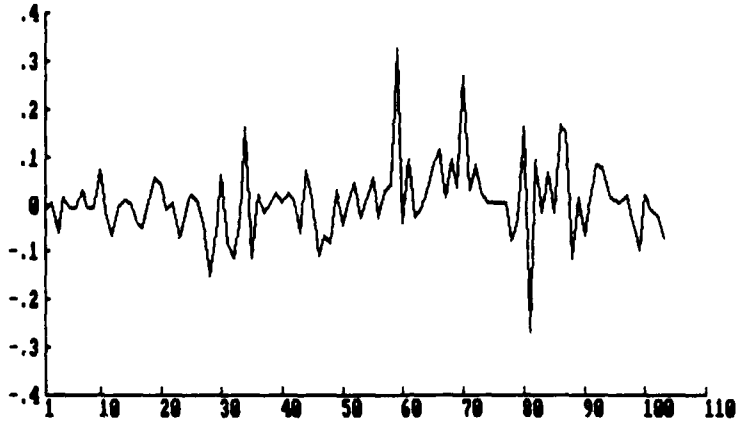

Fig. 34. *Observed weekly returns for H.F. Ahmanson and Company from July* 3, 1981 *to June* 30, 1983. *This is the first sample in the analysis.*

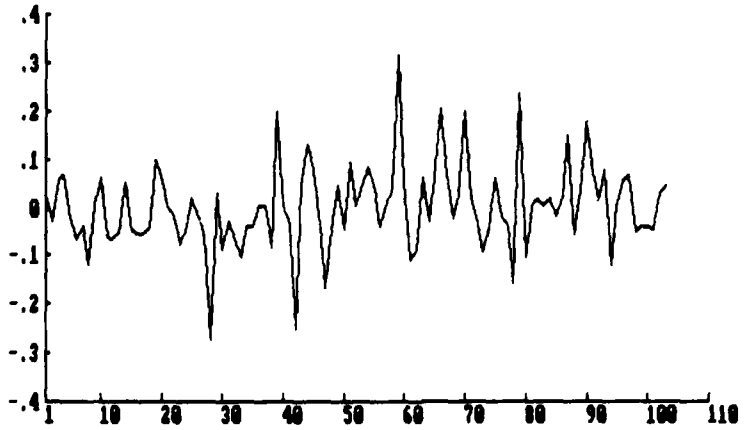

Fig. 35. *Observed weekly returns for Financial Corporation of Santa Barbara from July* 3, 1981 *to June* 30, 1983. *This is the second sample in the analysis.*

Table 16

Estimated small sample sizes of the subset chi-square test applied to components which are derived from correlated samples. The stated size of the test is 0.05.

	Number of Components	
Correlation	6	9
0.447	0.031	0.034
0.707	0.020	0.020
0.894	0.005	0.009
Bandwidth	0.3	0.2

insight into the behavior of the subset chi-square tests in this situation. The components are found and the subset chi-square test for the truncation points $M = 6, 9$ corresponding to bandwidths of $h = 0.3, 0.2$ is applied to them. The nominal size of the test is 5%. The indicator B_j is set to 1 if the test rejects and 0 if not. These indicators are then averaged over 1000 replications to estimate the size of the test. Table 16 gives the results. In each case, the estimated size is less than 5%. The higher the correlation, the lower is the estimated size. Since the test is conservative in this case and the reduction in its size for the level of correlation observed in the data is not extreme, the analysis will continue.

Table 17 presents the summary statistics for the two samples and the pooled sample. It is quite difficult to distinguish between the two based on these. Figures 36 through 38 present the identification quantile plots for the first, second, and pooled samples, respectively. These graphs are constructed following Parzen (1979) and described briefly here. A smoothed version of the sample quantile function for tne first sample is given by linearly interpolating the points

$$\hat{Q}(u) = X_{(j)} \text{ for } u = \frac{j - (1/2)}{m},\ j = 1, \ldots, m,$$

where $X_{(j)}$ is the j^{th} order statistic. The identification quantile function, QI(u), is defined as

$$\text{QI}(u) = (\hat{Q}(u) - \text{MQ})/\text{DQ},$$

Table 17

Sample statistics for the savings and loan data.

Statistic	First Sample	Second Sample	Pooled Sample
Median	0.0019	-0.0115	-0.0004
Twice Interquartile Range	0.1167	0.1971	0.1612
Maximum	0.3254	0.3151	0.3254
Minimum	-0.2667	-0.2733	-0.2733
Mean	0.0057	0.0000	0.0029
Variance	0.0060	0.0079	0.0069
Standard Deviation	0.0773	0.0891	0.0832

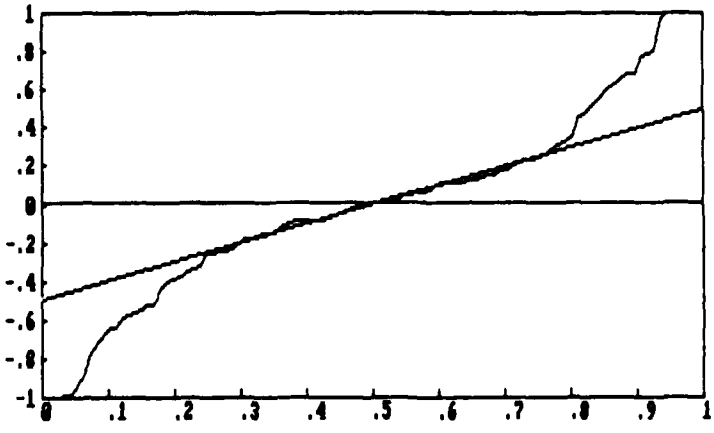

Fig. 36. *The identification quantile function of weekly returns for H.F. Ahmanson and Company.*

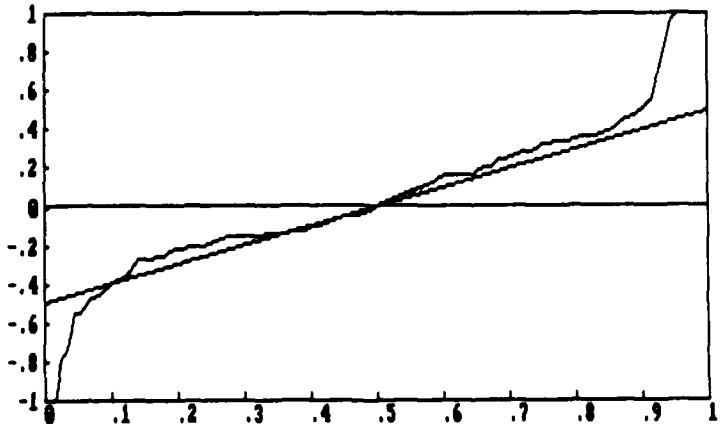

Fig. 37. *The identification quantile function of weekly returns for Financial Corporation of Santa Barbara.*

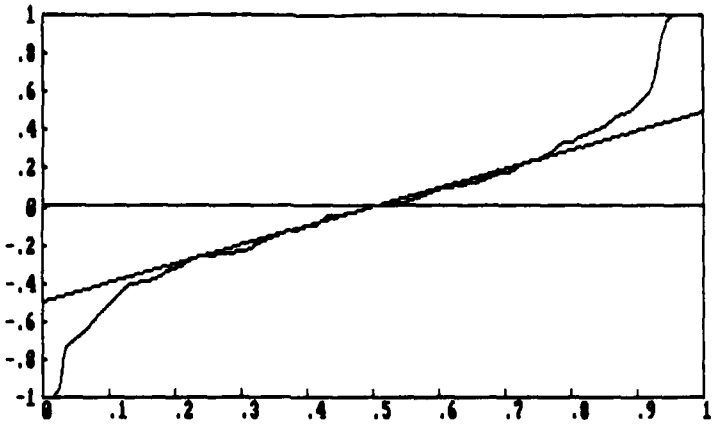

Fig. 38. *The identification quantile function of a pooling of weekly returns for H.F. Ahmanson and Company and Financial Corporation of Santa Barbara.*

where MQ is the sample median and $\mathrm{DQ} = 2[\hat{Q}(.75) - \hat{Q}(.25)]$ is the sample value of twice the interquartile range. The diagonal reference line in these figures is the identification quantile function of the uniform distribution. Normally distributed data will enter and exit near the corners of the box and have an inflection point at $u = 0.5$. By subtracting MQ and dividing by DQ, plots of the identification quantile function attempt to identify classes of distributions apart from location and scale.

Examining these figures, the data appear to have slightly longer than normal tails since they exit the boxes short of the corners. The three graphs appear quite similar. The identification quantile function of the first sample seems to follow the uniform reference line for a greater distance than the second sample. The fact that the identification quantile function of the pooled sample falls below the uniform reference line on about the range $u = 0.25$ to $u = 0.5$ where the others don't may indicate differences. However, there are difficulties in comparing plots such as these. It is hard to tell if the differences are really there or are due to random variation.

Figure 39 presents an overlay of the identification quantile functions for each sample. Here the differences for the left tail ($u < 0.5$) are brought into sharper

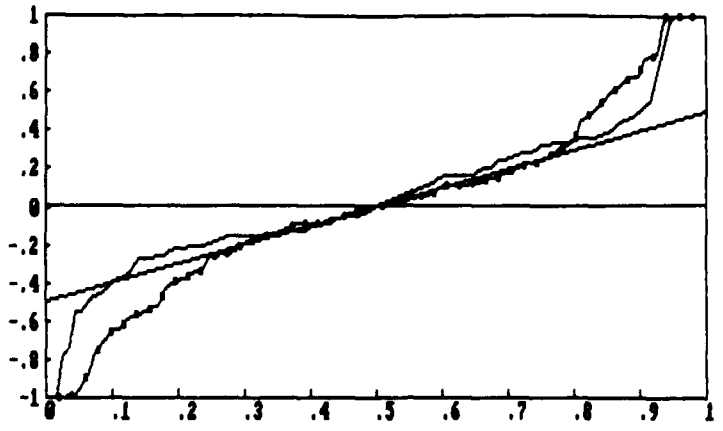

Fig. 39. *An overlay of the identification quantile functions of weekly returns for H.F. Ahmanson and Company (solid with blocks) and Financial Corporation of Santa Barbara (solid line).*

focus. Still, the question remains whether this difference is really there or is due to random variation. However, the purpose of these plots is to draw attention to such possibilities.

Figure 40 presents another type of plot often used to compare two populations. This is a QQ plot. Each box in the figure represents a pair $(X_{(k)}, Y_{(k)})$. The graph should be approximately linear with slope near 1 if the two populations are the same. The function appears quite linear with slope approximately equal to 1 except in the tail. Again, the question is how far away the pictured function must be from the ideal for the two populations to be declared different. Some work has been done in this area; see, for example, Aly and Bleuer (1986). Their work will not be pursued here.

Figure 41 gives the sample comparison distribution function, $D_N(u)$, for the data and a diagonal line for reference. Immediately apparent is the jump at $u = 0.5$. Although less apparent from the other graphs, it can be seen in Tables 14 and 15 that the data has repeats at a value of 0. There are some weeks that the stock price doesn't change. Repeat values violate the assumption that the distribution functions F and G are continuous. The analysis seems quite robust

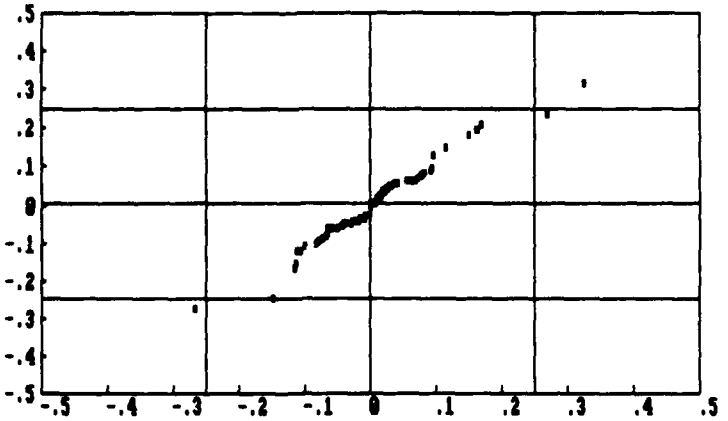

Fig. 40. *A QQ plot of the weekly returns of H.F. Ahmanson and Company (horizontal axis) and Financial Corporation of Santa Barbara (vertical axis).*

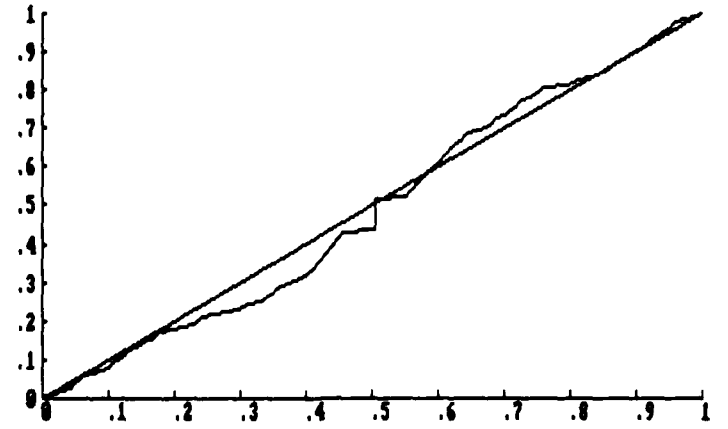

Fig. 41. *The sample comparison distribution function for the savings and loan return data.*

to this departure. The complete analysis has been repeated two different ways without a major change in results. The first was to add normal white noise with a very small variance to each series. This randomly breaks the ties. The results did not change. The second way was to conduct the analysis conditioned on the return being non-zero. Again, the analysis did not materially change. In the analysis conducted in this subsection, ties are resolved by midranking.

Returning to Figure 41, one sees some departure from uniformity but not enough for a clear rejection of H_o. More informative is Figure 42, which presents $\sqrt{N\lambda_{(N)}/(1-\lambda_{(N)}}[D_N(u)-u]$. Under H_o, this function converges weakly to a Brownian bridge process. The maximum absolute deviation of the function pictured in Figure 42 is not quite large enough for a Kolmogorov-Smirnov test to reject H_o. However, from the overall appearance of the graph, one might question whether there are enough zero crossings for this to be a sample path of a Brownian bridge process. The sample path seems somewhat deterministic. It is below 0 for $u < 0.5$ and mostly above 0 for $u > 0.5$. Table 18 presents the observed and critical values of the Cramér-von Mises, Anderson-Darling and Kolmogorov-Smirnov statistics. None rejects H_o at the 5% level.

Figure 43 presents diagnostics for the choice of bandwidth. The function pictured is given by linearly interpolating the points

$$\sqrt{\frac{N\lambda_{(N)}}{1-\lambda_{(N)}}}\left[\hat{D}_h(R_i/N)-\frac{i}{m+1}\right], \text{ for } i=1,\ldots,m, \tag{5.2.1}$$

where $\hat{D}_h(w)=\int_0^w \hat{d}_h(u)du$. If $\hat{D}_h(w)$ were the true comparison distribution function, the graphs should appear as a Brownian bridge process. Recall from Subsection 3.3.7 that the goal is to undersmooth the data slightly. Undersmoothing $\hat{d}_h(w)$ causes the deviations from 0 of the process defined in (5.2.1) to be too small. Referring to Figure 43, a bandwidth of 0.1 [Figure 43(a)] is clearly undersmoothing. Figure 43(b) ($h=0.2$) is also undersmoothed, but not as much: its deviations from 0 are larger. Figure 43(c) seems about the right amount of smoothing and Figure 43(d) appears to oversmooth. A bandwidth of $h=0.2$ is chosen.

Figure 44 gives the corresponding estimates of the comparison density func-

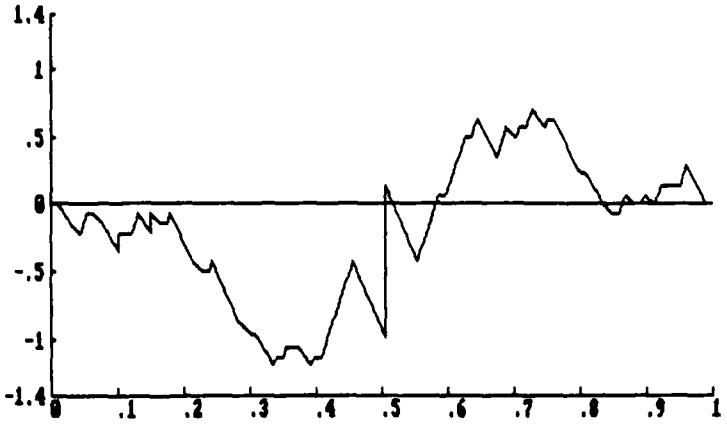

Fig. 42. *The sample null empirical comparison distribution process for the savings and loan return data. The function pictured is* $\sqrt{N}[D_N(u) - u]$.

Table 18

Two sample portmanteau statistics for the savings and loan data.

Statistic	Observed Value	5% Critical Value	1% Critical Value
Cramér-von Mises	0.288	0.460	0.740
Anderson-Darling	1.420	2.490	3.850
Kolmogorov-Smirnov	1.184	1.360	1.640
$\varphi^2_{0.2}$	9.304	8.931	12.400

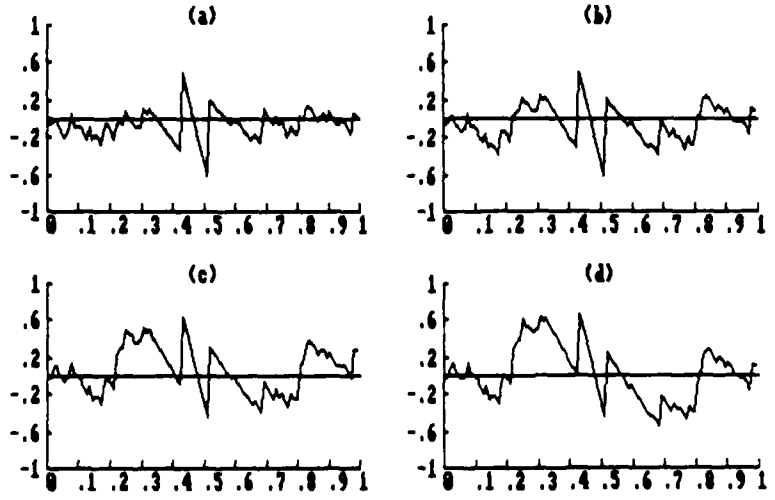

Fig. 43. *Sample paths of* $\sqrt{N}[\hat{D}_h(u) - u]$ *for the savings and loan return data. Figure* (a) *pictures the process for* $h = 0.1$*; Figure* (b) *for* $h = 0.2$*; Figure* (c) *for* $h = 0.3$*; and Figure* (d) *for* $h = 0.4$.

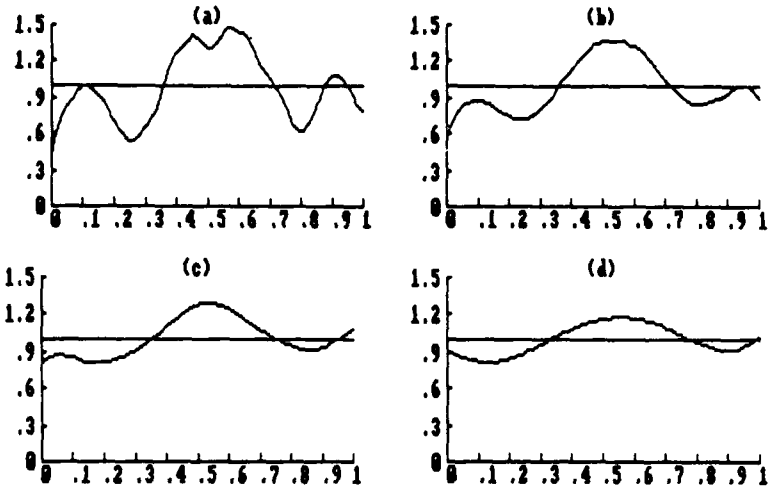

Fig. 44. *Boundary kernel estimates of the two sample comparison density function for the savings and loan return data. Figure (a) pictures the estimate for* $h = 0.1$*; Figure (b) for* $h = 0.2$*; Figure (c) for* $h = 0.3$*; and Figure (d) for* $h = 0.4$.

tion. These support the remarks made concerning the choice of bandwidth. Having chosen a bandwidth, the statistic $\varphi^2_{0.2}$ can be calculated. Its observed and critical values are presented in Table 18. The statistic rejects at the 5% level, but not the 1% level. This is the first solid evidence that the two samples come from distinct populations.

A truncation point of $M = 9$ is selected. This will include all eigenvalues above 0.01. Figure 45 gives the critical function, $C(k)$, for a test with size 5%. Its values are the blocks in the figure. The critical value for the next best subset for each size is given as the x's. Since $C(k)$ exceeds zero for some k, H_o is rejected. Table 19 gives the values of the squared components. The best subset includes the components 4, 8, 6, and 2, in that order. The value of $\chi^2_1(.95^{1/9})$ is 7.648. The independent tests method would include only the fourth component in the model. This observation supports earlier contentions about its behavior.

Figure 46 presents the boundary kernel estimate, $\hat{d}_{0.2}(w)$, and the orthogonal series estimate, $\hat{d}_{4,0.2}$, of the comparison density function. The blocks are the boundary kernel estimate and the solid line is the orthogonal series estimate.

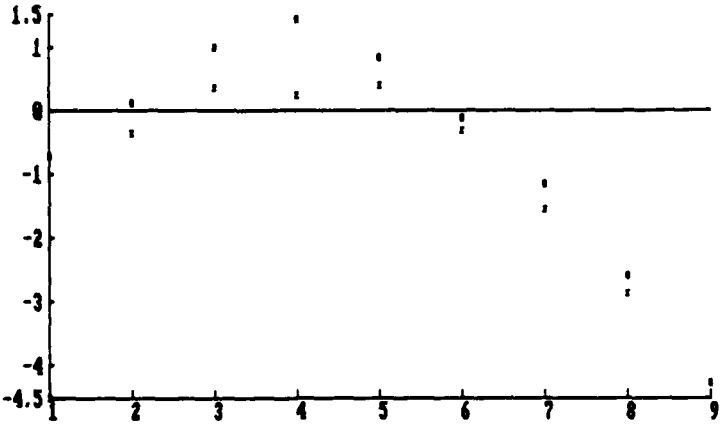

Fig. 45. *The criterion function, $C(k)$, for the savings and loan data. Also pictured are the criterion values for the next best subset.*

Table 19

The first nine squared components of the kernel density process ($h = 0.2$) for the savings and loan data.

Component Number	Squared Value
1	0.322
2	2.538
3	0.923
4	8.269
5	0.719
6	3.198
7	0.016
8	3.669
9	1.344

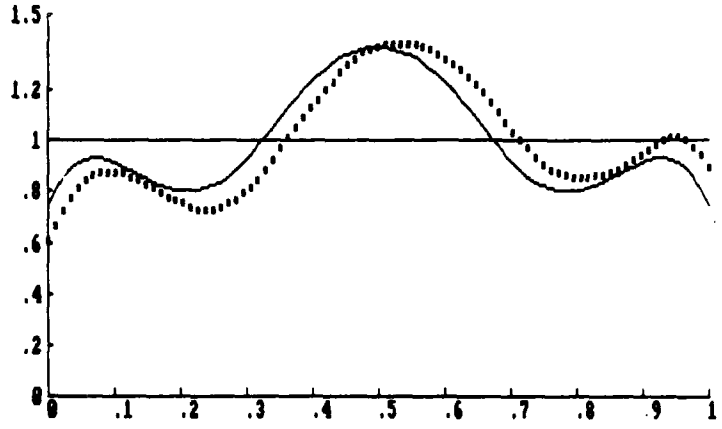

Fig. 46. *The boundary kernel estimate* ($h = 0.2$) *and the orthogonal series estimate* (*components* 4, 8, 6, *and* 2) *of the comparison density function for the savings and loan return data. The blocks are the boundary kernel estimate and the solid line is the orthogonal series estimate.*

The two have very much the same character. To interpret Figure 46 one starts by observing that $fQ^H(u) > gQ^H(u)$ for u values between about $u = 0.3$ and $u = 0.7$. The sample pooled quantile, $Q_N^H(u)$, appears largely symmetric about 0 (recall Figure 38). This implies that $f > g$ for values closer to 0 and that $f < g$ in the tails away from 0. This is not merely an artifact of the repeated values at 0. The first sample has 9 repeats at 0; the second sample 10.

From Figure 45 there appear to be two other models which should be examined. One has three components: 4, 8, and 6. The other has five components: 4, 8, 6, 2, and 9. Figure 47 presents the other two orthogonal series estimates. Figure 47(a) pictures the model with three components and Figure 47(b) pictures the model with five components. The former pictures the comparison density as rising back above 1 near both endpoints. The latter agrees substantially with the estimates pictured in Figure 46.

The Cramér-von Mises and Anderson-Darling statistics failed to reject H_o because the two samples have about the same location and scale. The $\varphi^2_{0.2}$ statistic detected a difference because it gives much greater weight to the fourth component than the other two statistics. The ratio of the fourth eigenvalue to

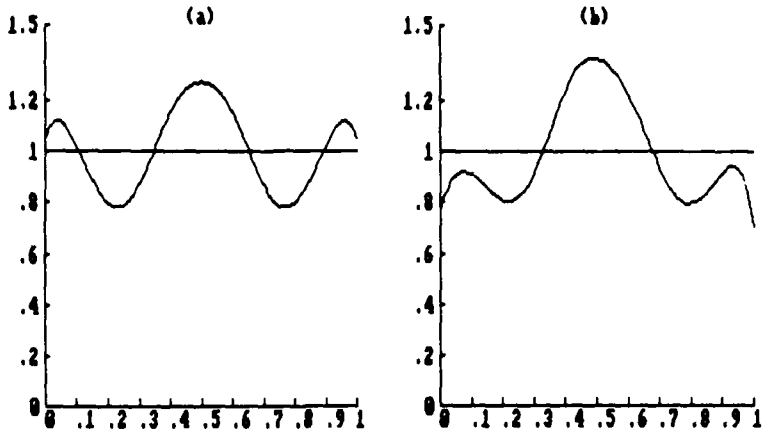

Fig. 47. *Alternate orthogonal series estimates of the comparison density function for the savings and loan return data. Figure* (*a*) *is based on a subset of size* 3 (*components* 4, 8, *and* 6) *and Figure* (*b*) *is based on a subset of size* 5 (*components* 4, 8, 6, 2 *and* 9).

the first for $\varphi^2_{0.2}$ is 0.49; for the Cramér-von Mises it is 0.0625; and for the Anderson-Darling it is 0.10.

The new methods suggested by this work were able to detect differences in the data that the standard portmanteau statistics (Kolmogorov-Smirnov, Cramér-von Mises, Anderson-Darling) could not. The procedure not only found differences, but was able to estimate the actual relation between the populations in a meaningful way. Interestingly, the φ^2_h statistic suggested by this work also detected a difference in the two populations.

5.3. The Behrens-Fisher Data

A simulated data set exhibiting the Behrens-Fisher problem is analyzed in this subsection. The Behrens-Fisher problem is to distinguish between two normal populations which differ in their mean and variance. The first sample is a random sample of size 30 drawn from the N(0,1) distribution. The second sample is a random sample of size 30 drawn from the N(1,3) distribution. The tests used in this example will clearly indicate that the null hypothesis is false. The value of such an example is to demonstrate the extra information that can be obtained

from the methods presented in this work.

The data are listed in Table 20 and presented graphically as Figure 48. Even from the figure it appears that the two samples are different. The second sample [Figure 48(b)] appears to be more variable if not possessing a greater mean. Table 21 gives the sample statistics for the two samples. These statistics also indicate likely differences in the two.

Figures 49 through 51 present the identification quantile plots for the first, second, and pooled samples, respectively. One would certainly not detect a difference based on these plots. Given the origin of the data, one would not expect to. The two populations are the same up to location and scale. These graphs remove the effect of location and scale. The pooled sample is bimodal, but this is not clear from the identification quantile function. Bimodality often causes the identification quantile to appear short tailed. The pooled quantile in Figure 51 doesn't appear to be short tailed. Figure 52 presents an overlay of the identification quantile plots for the two samples. The two appear very similar indeed.

Figure 53 presents the QQ plot of the two samples. The graph is somewhat deceiving because the horizontal and vertical axes are not in the same scale. The values above (0,0) do deviate substantially from the diagonal. The deviations below (0,0) are less severe. From this figure, one would strongly suspect that these two data sets are not from the same populations.

Table 22 gives the Cramér-von Mises, Anderson-Darling, Kolmogorov-Smirnov, and $\varphi^2_{0.3}$ statistics. Each rejects at the 5% level. The Kolmogorv-Smirnov also rejects at the 1% level. That H_o should be rejected is also clear from Figures 54 and 55. Figure 54 presents $D_N(u)$. It never falls below the reference diagonal line and the deviation from the diagonal is substantial. The process $\sqrt{N}[D_N(u)-u]$ drives the point home in Figure 55. The process certainly does not have the character of a Brownian bridge process.

Figure 56 pictures the process $\sqrt{N}[\hat{D}_h(u) - u]$. The bandwidth is selected in the same manner as before. Here a bandwidth of $h = 0.3$ is appropriate. A bandwidth of $h = 0.2$ might also be used. Figure 57 presents the corresponding

Table 20

White noise data exhibiting the Behrens-Fisher problem.

		First Sample			
0.243	0.258	1.082	-0.897	-0.713	1.486
-1.180	-3.147	0.722	1.108	2.048	0.764
0.062	1.800	0.684	0.462	-1.031	-1.560
-2.100	1.100	-0.250	-0.272	-0.432	0.117
0.047	-1.221	-0.800	-0.370	0.585	0.669
		Second Sample			
2.858	2.827	-0.260	1.202	-0.290	-1.202
-1.280	1.752	0.183	-0.731	-1.859	5.421
0.376	1.283	1.876	0.445	3.599	1.796
-1.752	-0.354	3.455	2.681	-2.880	1.599
-0.166	-0.796	2.270	2.387	2.278	1.524

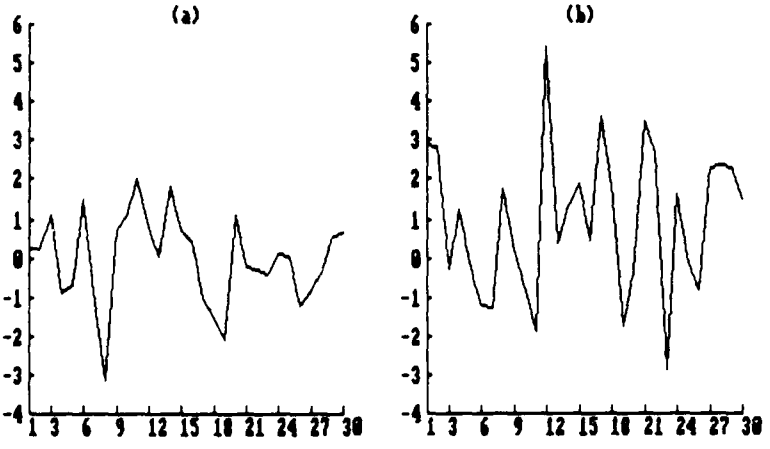

Fig. 48. *The data for the Behrens-Fisher problem. Figure (a) is the first sample and is a realization of a random sample from the* N(0,1) *distribution. Figure (b) is the second sample and is a realization of a random sample from the* N(1,3) *distribution.*

Table 21

Sample statistics for the Behrens-Fisher problem data.

Statistic	First Sample	Second Sample	Pooled Sample
Median	0.090	1.243	0.317
Twice Interquartile Range	3.044	5.265	4.567
Maximum	2.048	5.421	5.421
Minimum	-3.147	-2.880	-3.147
Mean	-0.025	0.941	0.458
Variance	1.347	3.607	2.672
Standard Deviation	1.161	1.899	1.635

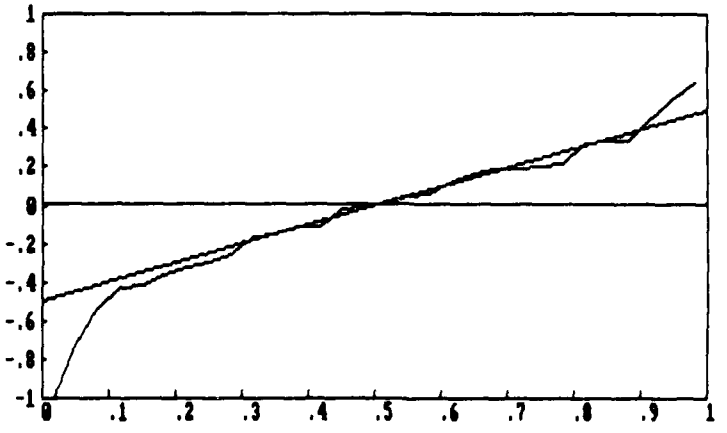

Fig. 49. *The identification quantile function of the first sample of the Behrens-Fisher problem.*

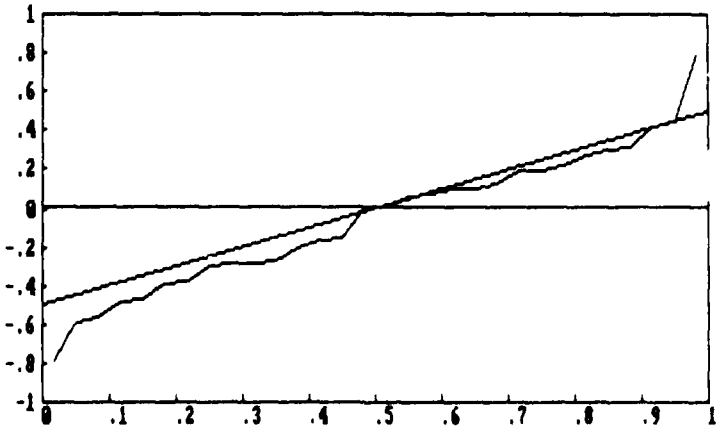

Fig. 50. *The identification quantile function of the second sample of the Behrens-Fisher problem.*

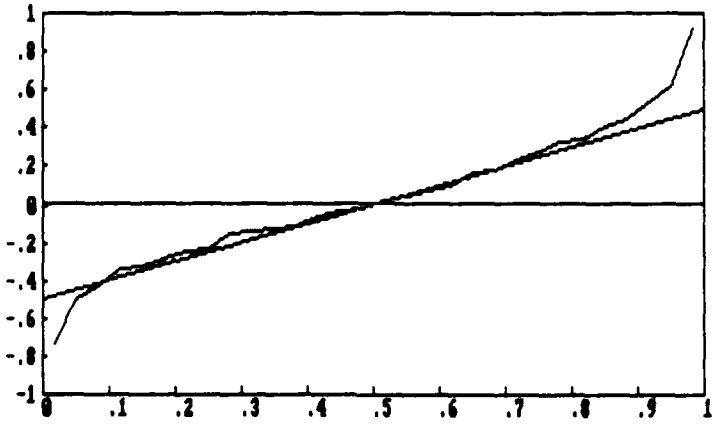

Fig. 51. *The identification quantile function of the pooled sample of the Behrens-Fisher problem.*

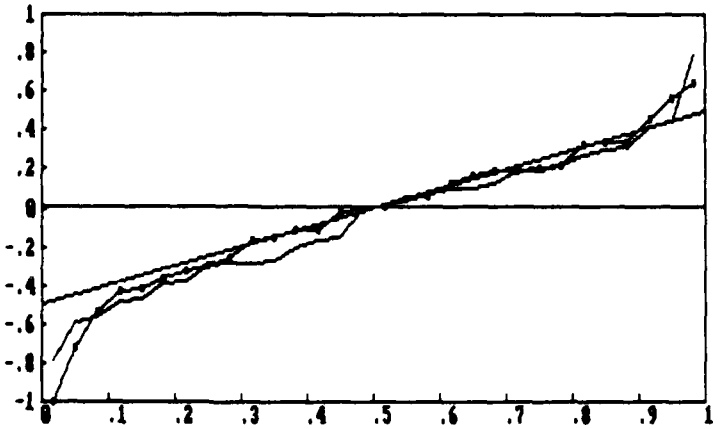

Fig. 52. *An overlay of the identification quantile functions of the two samples of the Behrens-Fisher problem. The solid line with blocks is the first sample and the solid line is the second sample.*

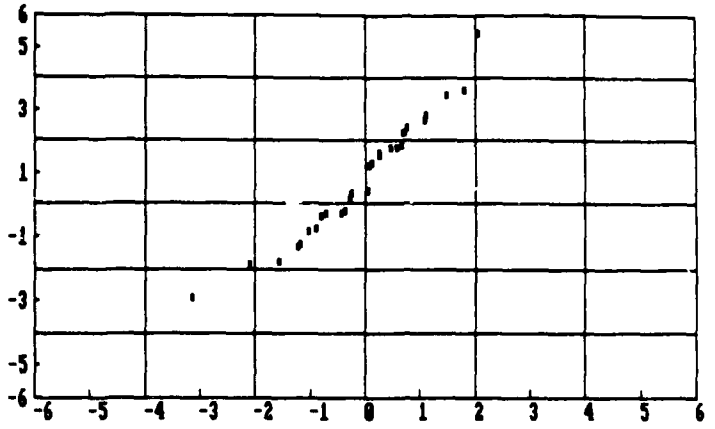

Fig. 53. *A QQ plot of the two samples of the Behrens-Fisher problem. The first sample is the horizontal axis and the second sample is the vertical axis.*

Table 22

Two sample portmanteau statistics for the Behrens-Fisher problem data.

Statistic	Observed Value	5% Critical Value	1% Critical Value
Cramér-von Mises	0.623	0.460	0.740
Anderson-Darling	3.532	2.490	3.850
Kolmogorov-Smirnov	1.678	1.360	1.640
$\varphi^2_{0.3}$	10.118	7.097	10.433

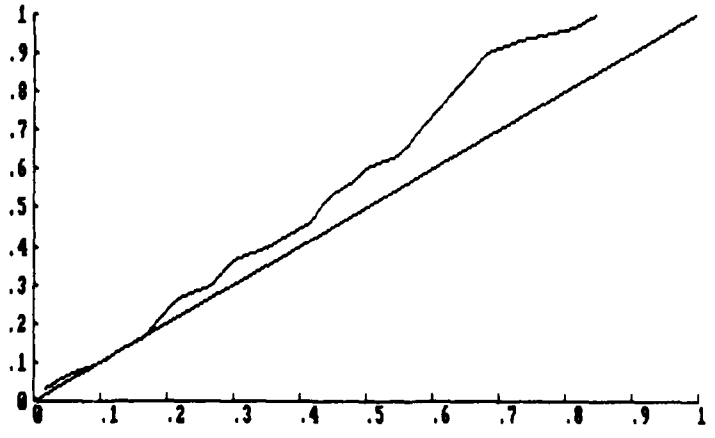

Fig. 54. *The sample comparison distribution function for the Behrens-Fisher problem.*

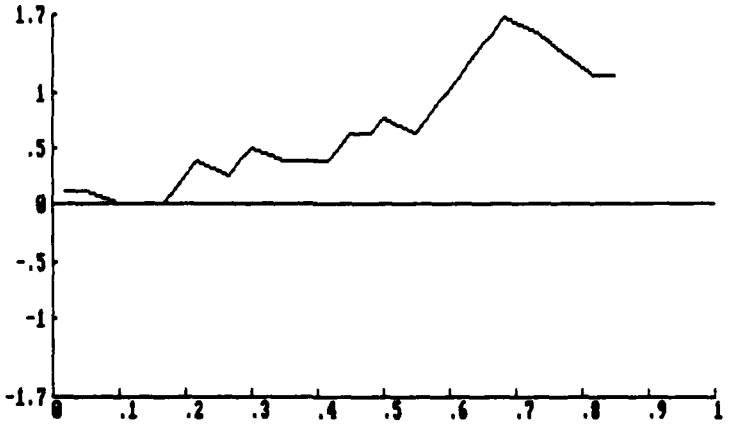

Fig. 55. *The sample null empirical comparison distribution process for the Behrens-Fisher problem.*

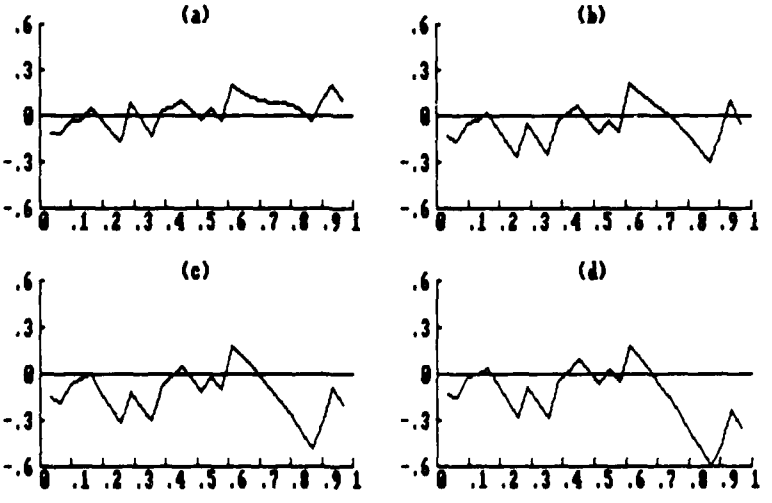

Fig. 56. *Sample paths of* $\sqrt{N}[\hat{D}_h(u) - u]$ *for the Behrens-Fisher problem. Figure* (a) *is constructed with* $h = 0.1$*; Figure* (b) *with* $h = 0.2$*; Figure* (c) *with* $h = 0.3$*; and Figure* (d) *with* $h = 0.4$.

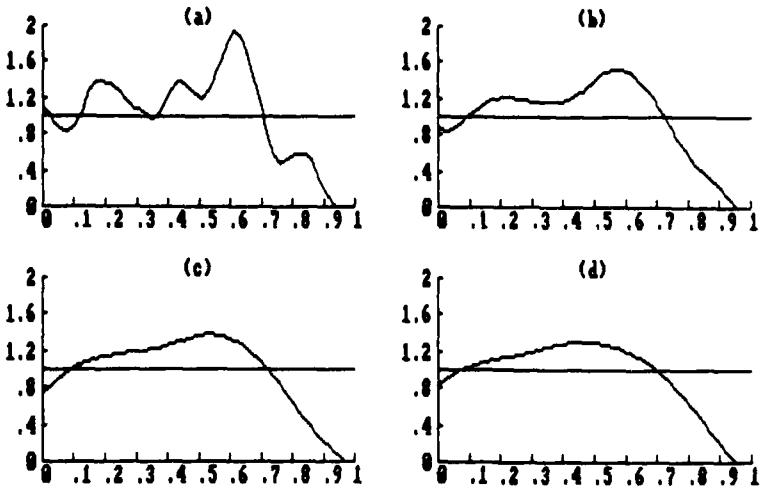

Fig. 57. *Boundary kernel estimates of the two sample comparison density function for the Behrens-Fisher problem. Figure (a) is constructed with* $h = 0.1$*; Figure (b) with* $h = 0.2$*; Figure (c) with* $h = 0.3$*; and Figure (d) with* $h = 0.4$.

comparison density estimates for these bandwidths. A truncation point of $M = 6$ is used with the bandwidth of $h = 0.3$. This truncation point includes all eigenvalues above 0.01.

Table 23 gives the squares of the components of the kernel density process for $h = 0.3$. The criterion function $C(k)$ for a size of 5% is given in Figure 58. The null hypothesis is rejected. A subset of size 2 is selected. The components in the most significant subset are 2 and 1. A subset of size 3 containing components 2, 1, and 5 might also be considered. The critical function is negative for each subset which yields the second largest value of $C(k)$. The value of $\chi^2_1(0.95^{1/6})$ is 6.922. Examining the squares of components in Table 23 one finds that the independent tests method would fail to reject H_o.

The boundary kernel and orthogonal series estimates of the comparison density function are presented in Figure 59. The orthogonal series estimate is somewhat smoother than the boundary kernel estimate. From the estimate, it appears the two samples differ mainly in scale. However, the first component is undeniably large. It is important that such numeric diagnostics accompany graphs to help direct the eye to important features.

Table 23

The first six squared components of the kernel density process ($h = 0.3$) for the Behrens-Fisher problem data.

Component Number	Squared Value
1	4.533
2	6.840
3	0.018
4	0.614
5	1.696
6	0.010

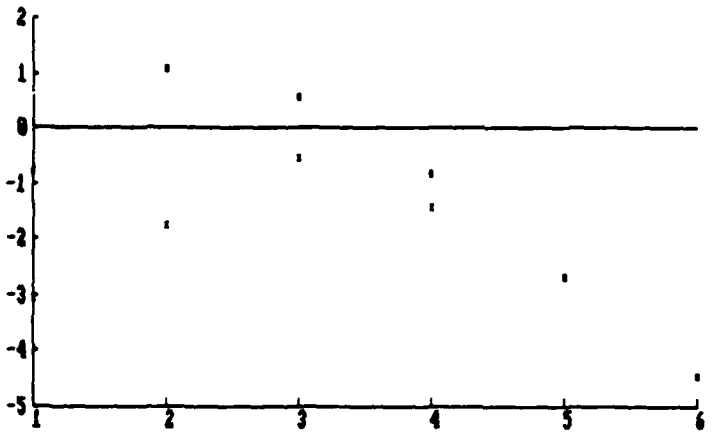

Fig. 58. *The criterion function, $C(k)$, for the Behrens-Fisher problem. Also pictured are the criterion values for the next best subset.*

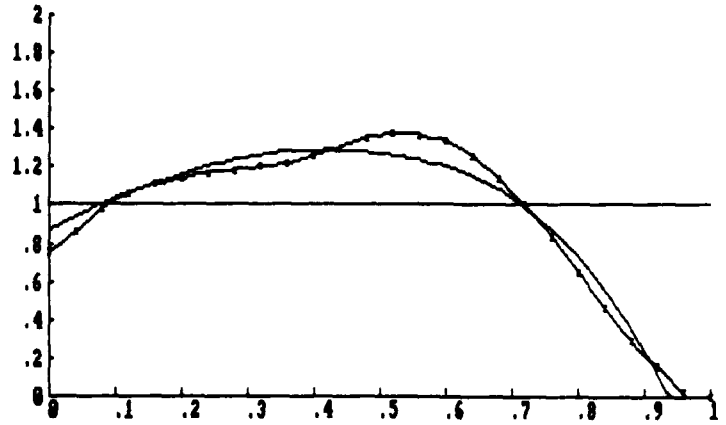

Fig. 59. *The boundary kernel estimate* ($h = 0.3$) *and orthogonal series estimate* (*components* 2 *and* 1), *of the comparison density function for the Behrens-Fisher problem. The line with blocks is the boundary kernel estimate and the solid line is the orthogonal series estimate.*

In this case it is possible to compare the estimated densities with the true comparison density. Figure 60 presents the true comparison density and the orthogonal series estimate. The estimate is excellent considering it was derived from two samples of size 30. In terms of estimating the region where $f > g$, the estimate misses on the left on an interval of length about 0.05 and on the right on an interval of length of only about 0.03. The square of the $\mathcal{L}_2$ distance between the estimated and true comparison density functions for the boundary kernel is 0.024 and 0.018 for the orthogonal series.

Each method (except the independent tests method applied to the components) rejected H_o. One can now judge the relative merit of each. The Kolmogorov-Smirnov, Cramér-von Mises, Anderson-Darling and φ_h^2 statistics give no indication of why they reject, only that they do. In combination with the identification quantile functions and sample statistics (MQ, DQ), one could discern that the two samples differ by location and scale factors. If the two differed by higher order components, these relationships would be much harder to identify in this manner. The estimate of the comparison density coupled with the components as diagnostics are as equally applicable to alternatives affecting

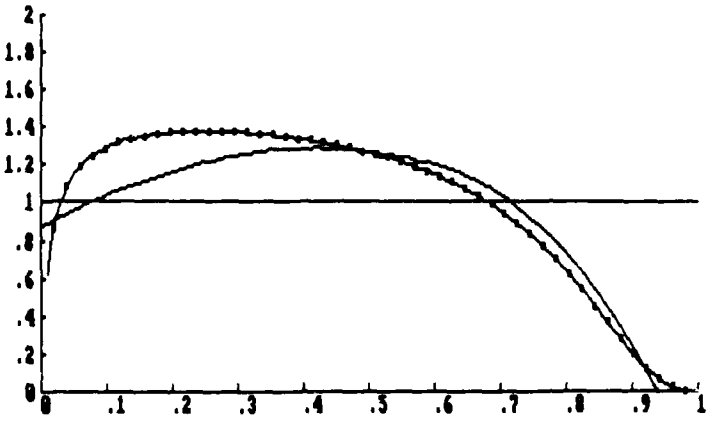

Fig. 60. *The orthogonal series estimate (components 2 and 1) and the actual comparison density function of the Behrens-Fisher problem. The line with blocks is the actual value and the solid line is the orthogonal series estimate.*

high and low order components.

The procedure has now been applied to two data sets. The first consisted of observed returns for two savings and loan institutions; the second was data simulated to exhibit the Behrens-Fisher problem. The first case exemplified an alternative which affects principally higher order components. The Kolmogorov-Smirnov, Cramér-von Mises, and Anderson-Darling tests all failed to detect a difference in the populations. The subset chi-square test applied to the components rejected H_o. The estimate of the comparison density gave an excellent graphical presentation of the relation of the two densities. The second data set exemplified an alternative for which every method rejects. Yet even here the estimate of the comparison density along with the components as diagnostics presented the relation of the populations in a clear and concise manner.

6. CONCLUSIONS

6.1. Conclusions

This work has sought to expand and refine the traditional analysis of two samples. The commonly used techniques were conceived in an era when computing facilities were a true constraint on what it was possible to do. Computing facilities no longer pose a constraint. Indeed, modern desktop personal computers and workstations are largely wasted on many traditional statistical methods. The analysis of two samples is one such area. A basic goal of this work has been to find a procedure more suited to the graphical and interactive computer environments now available.

A good deal more was sought in this work than just "gee whiz" type graphics and some number crunching. The desire has been to find a mode of analysis which provides a deeper understanding of the relation of the two populations under study. The philosophy has been that such a deeper understanding is possible now that numerically intensive methods are not ruled out and high quality graphics in real time are available.

Several desirable features that a procedure should possess were defined. It was desired to make minimal assumptions about the distribution functions of the two populations. A portmanteau test was desired to avoid specifying too closely the relation of the two distribution functions under alternatives. Similarly, a nonparametric test was required to avoid assuming a parametric family for the two distribution functions. Finally, it was desired to estimate the relation of F to G when H_o is rejected. Most existing two sample techniques fail to enlighten one in this regard.

Upon reviewing existing methodologies, it was seen that the comparison density is an important object in regards to several of these goals. The comparison density is uniform if and only if H_o is true. It is interpretable as the likelihood ratio of the density of the first sample to that of the pooled sample. An estimate of this density proves useful in two ways. First, it can be tested for uniformity

as a means of testing H_o. Second, it can serve as an estimate of the relation of the densities of the two samples.

Parzen (1983) introduced a natural estimate of the comparison distribution function. The comparison distribution function is simply the distribution function associated with the comparison density function. This estimator is called the sample comparison distribution function. The form of the sample comparison distribution function suggested strongly that a nonparametric density estimator be used to estimate the comparison density function. Results were to be derived based on the weak convergence of a centered and scaled version of the sample comparison distribution function. This centered and scaled process was called the empirical comparison distribution process.

The relevant nonparametric density estimation techniques were reviewed and it was decided to use the boundary kernel modification method of Gasser and Müller (1979). Several pointwise results were proved for the boundary kernel estimator of the comparison density function. Assuming the bandwidth shrinks to zero at an appropriate rate and several conditions on the kernel hold, it was proved that the boundary kernel estimate is asymptotically normal under H_o. Assuming a shrinking bandwidth, mild conditions on the kernel, and that the proportion of the total sample represented by the first sample doesn't change, the pointwise weak consistency of the boundary kernel estimate was proved. The boundary kernel estimate was also seen to be asymptotically invariant as to the choice of which sample is called the first.

A stochastic process called the kernel density process was introduced. It is a centered and scaled version of the boundary kernel estimate of the comparison density function. The weak convergence of the stochastic process was proved assuming mild conditions on the underlying kernel and a fixed bandwidth. Several rationales for a fixed bandwidth were given. Among these were: (1) under H_o the boundary kernel estimate is asymptotically unbiased for fixed bandwidths and (2) the fit of the small sample distribution to the asymptotic distribution is better if the latter is derived under a fixed bandwidth. The kernel density process forms the basis of testing the null hypothesis.

The results for the boundary kernel estimator of the comparison density function are significant in their own right. No comparable results exist in the literature. Processes such as the kernel density process are quite novel. This work marks a different mode of thinking about kernel density estimates in general.

A statistic, φ_h^2, was defined. It is a scaled version of the square of the $\mathcal{L}_2$ norm between the boundary kernel estimate of the comparison density function and the uniform density function. While investigating its limiting distribution, the idea of components of the kernel density process arose. The limiting distribution of φ_h^2 is an infinite weighted sum of squares of these components. The components were defined in detail. They were seen to be both generalized Fourier coefficients of an orthonormal expansion of the kernel density process and linear rank statistics. The orthonormal basis used is the set of eigenfunctions of the covariance kernel of the kernel density process under H_o. A numerical method for finding these eigenfunctions was presented. The space these functions span was seen to be of interest. This issue and its ramifications were investigated but not resolved.

The components were seen to be of more interest than the statistic which motivated them. It was proposed to base a test of the null hypothesis on the first M components and to give each equal weight in the test. This is in contrast to standard portmanteau statistics and φ_h^2 which employ all the components but successively downweight them.

Under H_o the components were proved to be asymptotically iid N(0,1). A method to test the components was needed. There are no optimal tests such as UMPU tests in this context. The two commonly used tests are the chi-square and the independent tests method. A new test was proposed instead of using one of these. The new test is called the subset chi-square test. It rejects H_o if and only if the sum of squares of some subset of the M components is found to be too large. Unlike the chi-square test, this test indicates which components are found to be significant. Unlike the independent tests method, the subset chi-square explicitly considers the components together and not just singly. As measured by power, the subset chi-square test was seen to be a good compromise between these other two.

The subset chi-square lends itself well to graphical display. Critical values for the test were found by simulation. By indicating components which are significant, the subset chi-square test suggested a subset orthogonal series estimator of the comparison density. The relation of the orthogonal series estimator to the boundary kernel estimator was investigated. The boundary kernel estimator was, itself, shown to be a damped orthogonal series estimator. The orthogonal series estimator suggested by the subset chi-square test simply includes or excludes particular frequencies.

The power of the subset chi-square, φ_k^2, the Cramér-von Mises, and Anderson-Darling statistics were compared in Section 4. These last two are commonly used portmanteau statistics. Powers for these tests were found for local location and scale alternatives for the normal and Cauchy distributions. Power functions were also calculated for what were termed Fourier alternatives. The two Fourier alternatives used stressed the third through sixth components. Location and scale alternatives significantly affected only the first through third components.

The weak convergence of the empirical comparison distribution process under local alternatives was proved. The power of the subset chi-square test applied to the components of the kernel density process was found by simulation. For the φ_k^2, Cramér-von Mises, and Anderson-Darling statistics, power functions were found by numerically inverting an approximation to their characteristic functions. A new method for this inversion based on the FFT was introduced. A theorem concerning the numerical consistency of the method was proved.

The subset chi-square was seen to perform very well. The traditional statistics performed better for location alternatives which affect mainly the first component. Their advantage disappeared for scale alternatives. For the Fourier alternatives, the standard statistics were found to be greatly lacking power in comparison to the subset chi-square test. These statistics may be consistent tests and the subset chi-square not, yet this seems little solace given their dismal performance as measured by power.

Lest one believe that alternatives seen in practice are only location or scale,

an example was given in Section 5. The data were observed weekly returns to two savings and loans over a 103 week period. All the standard portmanteau tests failed to reject H_o. The subset chi-square rejected H_o at the 5% level. The fourth component was dominant. Further, the estimate of the comparison density pictured the relation of the densities of the two populations in an understandable manner. A second example was also analyzed. This data was simulated to exhibit the Behrens-Fisher problem: the two samples were normal but with different means and variances. The differences were sufficiently large for the standard tests to detect them. Yet even in this case the procedure based on the comparison density gave unity and insight into the relation of the two samples.

In summary, the unified techniques based on the boundary kernel estimate of the comparison density achieve what was set out in Sections 1 and 2. The procedure is unified and self-contained. The test has good power against a wide range of alternatives. In fact, the breadth of this class is selected by the researcher. The procedure has many useful and informative graphical elements. The class of distributions to which it applies is very broad. When the test rejects H_o, it is also simultaneously selecting an orthogonal series estimate of the comparison density function. The orthogonal series estimate is intimately related to the boundary kernel estimate. The technique has been given a rigorous theoretical foundation. Its use will give the researcher an opportunity to more thoroughly understand his data and the information it contains.

6.2. Areas for Future Research

It is only natural to inquire where a piece of research will lead. Are there opportunities for expanding its scope? Are there other areas to which it applies? For the methods considered in this work, the answer is "yes" to both of these questions.

This research has concerned itself with the two sample problem. One could term it $k = 2$. It is only natural to ask about $k \geq 3$. This is the so-called k-sample problem and it should be an area rich for research. The basic approach would be to consider the ranks of each sample in a pooling of all k samples. One then has

$k - 1$ independent comparison densities to estimate. The choice of bandwidth across the samples and the method of testing the components would need to be considered in depth. There are certainly substantive issues to be addressed.

The alternative to increasing k above 2 is to decrease it to 1. This is known as the one sample problem. The one sample problem has just a single sample and tests a hypothesis of the form H_o: $F = F_o$, where F_o is some specified distribution function. The methods of this dissertation should apply almost wholesale to this problem. All one does is exchange the empirical comparison distribution process for the uniform empirical process. The rest of the analysis should apply almost directly.

In summary, there are very good prospects for expanding the methods of this work to related problems. The two most likely candidates for investigation are the k sample problem and the one sample problem.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. (1964). *Handbook of Mathematical Functions.* National Bureau of Standards, Applied Mathematics Series 55, Department of Commerce, Washington, D.C.

AHUÉS, M., D'ALMEIDA, F., CHATELIN, F., and TELIAS, M. (1982). Iterative Refinement Techniques for Eigenvalue Problem of Compact Integral Operators. In *Treatment of Integral Equations by Numerical Methods* (T. H. Baker and G. F. Miller, eds.) 373–385. Academic Press, New York.

AKAIKE, H. (1974). A New Look at Statistical Model Identification. *IEEE Trans. Auto. Cont.* **AC-19** 716-723.

ALY, E.-E. A. A. and BLEUER, S. (1986). Confidence Bands for Quantile-Quantile Plots. *Statist. and Decis.* **4** 205–225.

ALY, E.-E. A. A., CSÖRGŐ, M. and HORVÁTH, L. (1987). P–P Plots, Rank Processes and Chernoff-Savage Theorems. In *New Perspectives in Theoretical and Applied Statistics* (M. Puri, J. P. Vilaplann, and W. Wertz, eds.) 135–156. John Wiley & Sons, New York.

ANDERSON, T. W. and DARLING, D. A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *Ann. Math. Statist.* **23** 193–212.

BEAN, S. J. and TSOKOS, C. P. (1980). Developments in Nonparametric Density Estimation. *Int. Statist. Rev.* **48** 267–287.

BECKER, R. A., CHAMBERS, J. M. and WILKS, A. R. (1988). *The New S Language: A Programming Environment for Data Analysis and Graphics.* Wadsworth & Brooks/Cole, Pacific Grove, CA.

BERTRAND-RETALI, M. (1978). Convergence Uniforme d'un Estimateur de la Densité par la Méthode de Noyau. *Rev. Roumaine Math. Pures. Appl.* **23** 361–385.

BICKEL, P. J. and ROSENBLATT, M. (1973). On Some Global Measures of the Deviations of Density Function Estimates. *Ann. Statist.* **1** 1071–1095.

BICKEL, P. J. and DOKSUM, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden-Day, Inc., San Francisco.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* John Wiley & Sons, Inc., New York.

BILLINGSLEY, P. (1986). *Probability and Measure.* John Wiley & Sons, Inc., New York.

BOOS, D. (1986). Comparing K Populations with Linear Rank Statistics. *J. Amer. Statist. Assoc.* **81** 1018–1025.

BOWMAN, A. W. (1984). An Alternative Method of Cross-Validation for the Smoothing of Density Estimates. *Biometrika* **71** 353-360.

BOWMAN, A. W. (1985). A Comparative Study of some Kernel-Based Nonparametric Density Estimators. *J. Statist. Comp. Simul.* **21** 313–327.

BROWN, S. and WARNER, J. B. (1980). Measuring Security Price Performance. *J. Finan. Econ.* **7** 205–258.

CARMICHAEL, J. P. (1984). Consistency of an Autoregressive Density Estimator. *Math. Operations-forsch. Statist. Ser. Statist.* **15** 383–387.

CENCOV, N. N. (1962). Evaluation of an Unknown Distribution Density from Observations. *Sov. Math.* **3** 1559–1562.

CHERNOFF, H. and SAVAGE, I. R. (1958). Asymptotic Normality and Efficiency of Certain Nonparametric Test Statistics. *Ann. Math. Statist.* **29** 972-994.

CHOW, Y.-S., GEMAN, S. and WU, L.-D. (1983). Consistent Cross-Validated Density Estimation. *Ann. Statist.* **11** 25–38.

DANN, L. V. and JAMES, C. M. (1982). An Analysis of the Impact of Deposit Rate Ceilings on the Market Values of Thrift Institutions. *J. Finan.* **37** 1259–1275.

DIGGLE, P. J. and HALL, P. (1986). The Selection of Terms in an Orthogonal Series Density Estimator. *J. Amer. Statist. Assoc.* **81** 230–233.

DUIN, R. P. W. (1976). On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions. *IEEE Trans. Comput.* **C-25** 1175–1179.

DURBIN, J. and KNOTT, M. (1972). Components of Cramér-von Mises Statistics. I. *J. Roy. Statist. Soc. B* **34** 290–307.

EUBANK, R. L., LARICCIA, V. N. and ROSENSTEIN, R. B. (1987). Test Statistics Derived as Components of Pearson's Phi-Squared Distance Measure. *J. Amer. Statist. Assoc.* **82** 816–825.

FAMA, E. F. (1976). *Foundations of Finance.* Basic Books, Inc., New York.

FISHER, R. A. (1948). *Statistical Methods for Research Workers.* Oliver and Boyd, London.

FISZ, M. (1960). On a Result by M. Rosenblatt Concerning the von Mises-Smirnov Test. *Ann. Math. Statist.* **31** 427–429.

FURNIVAL, G. M. (1971). All Possible Regressions with Less Computation. *Technometrics* **13** 403–408.

FURNIVAL, G. M. and WILSON, R. W. (1974). Regressions by Leaps and Bounds. *Technometrics* **16** 499–511.

GAENSSLER, P. (1983). *Empirical Processes.* Institute of Mathematical Statistics Lecture Notes-Monograph Series, Hayward, CA.

GASSER, T. and MÜLLER, H.-G. (1979). Kernel Estimation of Regression Functions. In *Smoothing Techniques for Curve Estimation, Proceedings, Heidelberg 1979* (T. Gasser and M. Rosenblatt, eds.) 23–68. Springer-Verlag, Berlin.

GOOD, I. J. and GASKINS, R. A. (1971). Nonparametric Roughness Penalties for Probability Densities. *Biometrika* **58** 255–277.

HALL, P. (1981). On Trigonometric Series Estimates of Densities. *Ann. Statist.* **9** 683–685.

HALL, P. (1983a). Large Sample Optimality of Least Squares Cross-Validation in Density Estimation. *Ann. Statist.* **11** 1156–1174.

HALL, P. (1983b). Measuring Efficiency of Trigonometric Series Estimates of a Density. *J. Multivariate Anal.* **13** 234–256.

HARPAZ, A. (1985). Stationary Times Series, Quantile Functions, Nonparametric Inference and Rank Transform Spectrum. Department of Statistics, Texas A&M University, Technical Report A-31.

HART, J. D. (1985). On the Choice of Truncation Point in Fourier Series Density Estimation. *J. Statist. Comp. Simul.* **21** 95–116.

HART, J. D. (1988). An ARMA Type Probability Density Estimator. *Ann. Statist* **16** 842–855.

HOCKING, R. R. (1985). *The Analysis of Linear Models.* Brooks/Cole Publishing Co., Monterey, CA.

IMHOF, J. P. (1961). Computing the Distribution of Quadratic Forms in Normal Variables. *Biometrika* **48** 419–426.

JONES, M. C. and LOTWICK, H. W. (1984). A Remark on Algorithm AS 176 Kernel Density Estimation Using the Fast Fourier Transform. Remark AS R50. *Appl. Statist.* **33** 120–122.

KANNAN, D. (1979). *An Introduction to Stochastic Processes.* North Holland, New York.

KELLER, G., WARRACK, B. and BARTEL, H. (1988). *Statistics for Management and Economics: A Systematic Approach.* Wadsworth Publishing Co., Belmont, CA.

KENDALL, M. and STUART, A. (1979). *The Advanced Theory of Statistics, Volume 2, Inference and Relationship.* Macmillan Publishing Co., Inc., New York.

KIEFER, J. (1959). K-Sample Analogues of the Kolmogorov-Smirnov and Cramér-von Mises Tests. *Ann. Math. Statist.* **30** 420–447.

KRONMAL, R. and TARTER, M. (1968). The Estimation of Probability Densities and Cumulatives by Fourier Series Methods. *J. Amer. Statist. Assoc.* **63** 925–952.

LEHMANN, E. L. (1951). Consistency and Unbiasedness of Certain Nonparametric Tests. *Ann. Math. Statist.* **22** 165–179.

LEHMANN, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, Inc., San Francisco.

LUKAS, M. A. (1980). Regularization. In *The Application and Numerical Solution of Integral Equations* (R. S. Anderssen, F. R. de Hoog and M. A. Lukas, eds.) 151–182. Sijthoff and Noordhoof, Germantown, MD.

MARTI, J. T. (1982). On a Regularization Method for Fredholm Equations of the First Kind Using Sobolev Spaces. In *Treatment of Integral Equations by Numerical Methods* (C. T. H. Baker and G. F. Miller, eds.) 59–66. Academic Press, New York.

MONTGOMERY, D. C. (1984). *Design and Analysis of Experiments.* John Wiley & Sons, Inc., New York.

MOOD, A. M., GRAYBILL, F. A., and BOES, D. C. (1974). *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., New York.

NADARYA, E. A. (1965). On Nonparametric Estimates of Density Functions and Regression Curves. *Theor. Prob. Appl.* **10** 186–190.

NEWTON, H. J. (1988). *Timeslab: A Time Series Analysis Laboratory.* Wadsworth & Brooks/Cole, Pacific Grove, CA.

PARZEN, E. (1962a). On Estimation of a Probability Density Function and Mode. *Ann. Math. Statist.* **33** 1065–1076.

PARZEN, E. (1962b). *Stochastic Processes.* Holden-Day, Inc., San Francisco.

PARZEN, E. (1977). Multiple Time Series: Determining the Order of Approximating Autoregressive Schemes. In *Multivariate Analysis-IV* (P. Khrishnaiah, ed.) 283–295. North Holland, Amsterdam.

PARZEN, E. (1979). Nonparametric Statistical Data Modeling. *J. Amer. Statist. Assoc.* **74** 105–131.

PARZEN, E. (1983). Fun.Stat Quantile Approach to Two Sample Statistical Data Analysis. Institute of Statistics, Texas A&M University, Technical Report A-21.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A., and VETTERLING, W. T. (1986). *Numerical Recipes: the Art of Scientific Computing.* Cambridge University Press, New York.

PRIHODA, T. J. (1981). A Generalized Approach to the Two Sample Problem: the Quantile Approach. Institute of Statistics, Texas A&M University, Technical Report B-5.

PYKE, R. and SHORACK, G. (1968). Weak Convergence of a Two Sample Empirical Process and a New Approach to Chernoff-Savage Theorems. *Ann. Math. Statist.* **39** 755–771.

RANDLES, R. H. and HOGG, R. V. (1971). Certain Uncorrelated and Independent Rank Statistics. *J. Amer. Statist. Assoc.* **66** 569–574.

RANDLES, R. H. and WOLFE, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics.* John Wiley & Sons, New York.

RICE, J. (1984). Boundary Modification for Kernel Regression. *Commun. Statist.-Theor. Meth.* **13** 893–900.

ROSENBLATT, M. (1952). Limit Theorems Associated with Variants of the von Mises Statistic. *Ann. Math. Statist.* **23** 617–623.

ROSENBLATT, M. (1956). Remarks on some Nonparametric Estimates of a Density Function. *Ann. Math. Statist.* **27** 832–837.

RUDEMO, M. (1982). Empirical Choice of Density Estimators. *Scand. J. Statist.* **9** 65–78.

RUYMGAART, F. (1988). Unpublished lecture notes. Department of Statistics, Texas A&M University (private collection W. Alexander).

SCHUCANY, W. R. and SOMMERS, J. P. (1977). Improvement of Kernel Type Density Estimators. *J. Amer. Statist. Assoc.* **72** 420–423.

SCHUSTER, E. F. (1985). Incorporating Support Constraints into Nonparametric Estimates of Densities. *Commun. Statist.-Theor. Meth.* **14** 1123–1136.

SCOTT, D. W., TAPIA, R. A., and THOMPSON, J. R. (1977). Kernel Density Estimation Revisited. *Nonlin. Anal.* **1** 339-372.

SCOTT, D. W. (1979). On Optimal and Data-Based Histograms. *Biometrika* **66** 605–610.

SCOTT, D. W., and TERRELL, G. R. (1987). Biased and Unbiased Cross-Validation in Density Estimation. *J. Amer. Statist. Assoc.* **82** 1131–1146.

SEHEULT, A. H., and QUESENBERRY, C. P. (1971). On Unbiased Estimation of Density Functions. *Ann. Math. Statist.* **42** 1434–1438.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York.

SHORACK, G. and WELLNER, J. (1986). *Empirical Processes with Applications to Statistics.* John Wiley & Sons, New York.

SILVERMAN, B. W. (1978). Weak and Strong Consistency of the Kernel Estimate of a Density and its Derivatives. *Ann. Statist.* **6** 177–184.

SILVERMAN, B. W. (1982). Kernel Density Estimation using the Fast Fourier Transform. Statistical Algorithm AS 176. *Appl. Statist.* **31** 93–97.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

SILVEY, S.D. (1975). *Statistical Inference.* Chapman and Hall, London.

SLEPIAN, D. (1958). Fluctuations in Random Noise Power. *The Bell System Tech. J.* **37** 163–184.

STONE, C. J. (1984). An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates. *Ann. Statist.* **12** 1285–1297.

STROMBERG, K. R. (1981). *An Introduction to Classical Real Analysis.* Wadsworth International Group, Belmont, CA.

TAPIA, R. A. and THOMPSON, J. R. (1978). *Nonparametric Probability Density Estimation.* The Johns Hopkins University Press, Baltimore.

TARTER, M. E. and KRONMAL, R. A. (1976). An Introduction to the Implementation and Theory of Nonparametric Density Estimation. *Amer. Statist.* **30** 105–112.

TITTERINGTON, D. M. (1985). Common Structure of Smoothing Techniques in Statistics. *Int. Statist. Rev.* **53** 141–170.

VAN ZWET, W. R. (1983). Rank and Order Statistics. In *Recent Advances in Statistics: Papers in Honor of Herman Chernoff on his Sixtieth Birthday* (M. H. Rizvi, J. Rustagi, and D. Siegmund, eds.) 407–422. Academic Press, New York.

WAHBA, G. (1975). Optimal Convergence Properties of Variable Knot, Kernel, and Orthogonal Series Methods of Density Estimation. *Ann. Statist.* **3** 15–29.

WAHBA, G. (1981). Data Based Optimal Smoothing of Orthogonal Series Density Estimates. *Ann. Statist.* **9** 146–156.

WATSON, G. S. (1969). Density Estimation by Orthogonal Series. *Ann. Math. Statist.* **40** 1496–1498.

WEGMAN, E. J. (1972a). Nonparametric Probability Density Estimation: I. A Summary of Available Methods. *Technometrics* **14** 533-546.

WEGMAN, E. J. (1972b). Nonparametric Probability Density Estimation: II. A Comparison of Density Estimation Methods. *J. Statist. Comp. Simul.* **1** 225–245.

WERTZ, W. (1978). *Statistical Density Estimation: A Survey.* Vandenhoeck and Ruprecht, Göttingen.

WOODROOFE, M. (1970). On Choosing a Delta Sequence. *Ann. Math. Statist.* **41** 1665–1671.

APPENDIX A

GLOSSARY OF NOTATION

Appendix A is a glossary of the notation used in this work. In general, if a function is subscripted by (N), the function is not random. The dependence on N is through $\lambda_{(N)}$, the fraction of the pooled sample represented by the first sample. For example, the function $D_{(N)}(u)$ is not random. If a function is subscripted by N, then that function is random. For example, the function $D_N(u)$ is random. Following is a list of the notation used in this work along with a brief explanation.

- The first sample is $X_1, \ldots, X_m$ which is iid with distribution function F, density function f, and quantile function Q^F.
- The second sample is $Y_1, \ldots, Y_n$ which is iid with distribution function G, density function g, and quantile function Q^G.

or

- The second sample is $Y_{n1}, \ldots, Y_{nn}$ which is iid with distribution function $G_{(n)}$, density function $g_{(n)}$, and quantile function $Q^G_{(n)}$ for local alternatives.
- The first sample has m observations.
- The second sample has n observations.
- The pooled sample has $N = n + m$ observations.
- The empirical distribution functions are:

F_m for the first sample,

G_n for the second sample,

H_N for the pooled sample.

- The empirical quantile functions are:

Q^F_m for the first sample,

Q^G_n for the second sample,

Q^H_N for the pooled sample.

• The ratio of the size of the first sample to the size of the pooled sample is

$$\lambda_{(N)} = m/N.$$

This fraction satisfies one of the two following conditions:

$$\lambda_{(N)} \to \lambda_0 \text{ as } m \wedge n \to \infty,\ 0 < \lambda_0 < 1,$$
$$\lambda_{(N)} = \lambda_0,\ 0 < \lambda_0 < 1.$$

• The true or population distribution function of the pooled sample is

$$H_{(N)}(x) = \lambda_{(N)}F(x) + (1 - \lambda_{(N)})G(x),$$

or

$$H_o(x) = \lambda_0 F(x) + (1 - \lambda_0)G(x),$$

or

$$H_{(N)}(x) = \lambda_{(N)}F(x) + (1 - \lambda_{(N)})G_{(n)}(x),$$

for fixed m and n, as $m \wedge n \to \infty$ and for local alternatives, respectively.

• The true or population quantile function of the pooled sample is

$$Q^H_{(N)}(u) = H^{-1}_{(N)}(u),$$

or

$$Q^H_o(u) = H^{-1}_o(u),$$

for fixed m and n and as $m \wedge n \to \infty$, respectively.

• The sample comparison distribution function is

$$D_N(u) = (H_n Q^F_m)^{-1}(u).$$

• The population comparison distribution function is

$$D_{(N)}(u) = FQ^H_{(N)}(u),$$
$$D_o(u) = FQ^H_o(u),$$
$$D_\lambda(u) = FQ^H_\lambda(u),$$

for fixed m and n, as $m \wedge n \to \infty$, and for $\lambda_{(N)}$ equal to arbitrary λ, respectively.

- The population comparison density function is

$$d_{(N)}(u) = D'_{(N)}(u),$$
$$d_o(u) = D'_o(u),$$
$$d_\lambda(u) = D'_\lambda(u),$$

for fixed m and n, as $m \wedge n \to \infty$, and for $\lambda_{(N)}$ equal to arbitrary λ, respectively.

- The comparison distribution empirical process is

$$\mathrm{CD}_N(u) = \sqrt{N}[D_N(u) - D_{(N)}(u)].$$

- The null comparison distribution empirical process is

$$\mathrm{CDo}_N(u) = \sqrt{N}[D_N(u) - u],$$

which equals CD_N under H_o.

- The limiting process of $\mathrm{CD}_N(u)$ is

$$L(u).$$

- The boundary kernel estimate of the comparison density is

$$\hat{d}_h(w) = \frac{1}{h}\int_0^1 K_s([w-u]/h)dD_N(u).$$

- The sample kernel density process is

$$\mathrm{KDP}_{N,h}(w) = \frac{1}{h}\int_0^1 K_s([w-u]/h)d\mathrm{CD}_N(u).$$

- The null sample kernel density process is

$$\mathrm{KDPo}_{N,h}(w) = \frac{1}{h}\int_0^1 K_s([w-u]/h)d\mathrm{CDo}_N(u)$$
$$= \sqrt{N}[\hat{d}_h(w) - 1].$$

- The limiting process of the sample kernel density process is

$$\mathrm{KDP}_h(w) = \frac{1}{h^2}\int_0^1 K'_s([w-u]/h)L(u)du.$$

- A normalized estimate of Pearson's phi-squared statistic is

$$\hat{\varphi}^2_{N,h} = \frac{N\lambda_{(N)}}{1-\lambda_{(N)}}\int_0^1 [\hat{d}_h(w) - 1]^2 dw.$$

It converges in distribution to

$$\varphi^2_h.$$

APPENDIX B

PROOFS OF THEOREMS AND LEMMAS

Proof of Theorem 3.2.1

Before starting on the proof of Theorem 3.2.1, a quote from Van Zwet (1983) seems appropriate. He discusses the proof of the original Chernoff-Savage theorem and subsequent developments:

"It (Chernoff and Savage's proof) struck terror into the hearts of graduate students at the time because of—what was then considered—its extreme technicality; in order to approximate the rank statistic by a sum of independent random variables no fewer than six remainder terms were shown to tend to zero, each for its own particular reason. Unfortunately, the number of such remainder terms has increased monotonically over the years and nowadays authors in this area appear to need at least fifteen."

Start by writing $\sqrt{Nh}[\hat{d}_h(w) - 1]$ as

$$\sqrt{Nh}\left[\int_{-\infty}^{\infty} \frac{1}{h} K([w - H_N(t)]/h) dF_m(t) - 1\right].$$

Let $t = Q^F(u)$ and perform a change of variable to arrive at the following equivalent expression:

$$\sqrt{Nh}\left[\int_0^1 \frac{1}{h} K([w - H_N Q^F(u)]/h) dF_m Q^F(u) - 1\right].$$

These statements hold true for each $w \in (0,1)$ for h sufficiently small. Define the uniform empirical distribution function for the first sample as $\Gamma_m^F(u) = F_m Q^F(u)$ and $\Gamma_N^H(u) = H_N Q^F(u)$, for the pooled sample. Under H_o, this last process is also a uniform empirical distribution function. Substituting these quantities in the above integral, one arrives at

$$\begin{aligned} &\sqrt{Nh}[\hat{d}_h(w) - 1] \\ &= \sqrt{Nh}\left[\int_0^1 \frac{1}{h} K([w - \Gamma_N^H(s)]/h) d\Gamma_m^F(s) - 1\right]. \end{aligned} \tag{B.1}$$

The mean value theorem states that for each $s \in [0,1]$ there exists a point $t_N(s)$ between $\Gamma_N^H(s)$ and s such that

$$K([w-\Gamma_N^H(s)]/h) = K([w-s]/h) - \frac{1}{h}K'([w-t_N(s)]/h)[\Gamma_N^H(s)-s]$$

The result of the application of the mean value theorem can now be substituted into (B.1). This yields

$$\sqrt{Nh}\left\{\frac{1}{h}\int_0^1 \left[K([w-s]/h) - \frac{1}{h}K'([w-t_N(s)]/h)[\Gamma_N^H(s)-s]\right] d\Gamma_m^F(s) - 1\right\}$$
$$= A_{1N} - A_{2N} - B_{1N} - B_{2N},$$

where

$$A_{1N} = \sqrt{Nh}\left\{\frac{1}{h}\int_0^1 K([w-s]/h)d\Gamma_m^F(s) - 1\right\},$$
$$A_{2N} = \frac{1}{h^{1.5}}\int_0^1 K'([w-s]/h)U_N^H(s)ds,$$
$$B_{1N} = \frac{1}{h^{1.5}}\int_0^1 \left[K'([w-t_N(s)]/h) - K'([w-s]/h)\right] U_N^H(s)ds,$$
$$B_{2N} = \frac{1}{h^{1.5}}\int_0^1 K'([w-t_N(s)]/h)U_N^H(s)d[\Gamma_m^F(s)-s],$$

and

$$U_N^H(s) = \sqrt{N}[\Gamma_N^H(s)-s]$$

is the uniform empirical process for the pooled sample. The first order terms are A_{1N} and A_{2N}. They will be shown to converge in distribution. The second order terms are B_{1N} and B_{2N}. They will be shown to converge in probability to zero. Start with the first order terms. One can rewrite A_{1N} as

$$A_{1N} = \sqrt{Nh}\left\{\frac{1}{mh}\sum_{i=1}^{m} K([w-U_i]/h) - 1\right\}$$

where $U_i = F(X_i)$ is uniformly distributed. One can rewrite A_{2N} as

$$A_{2N} = \frac{1}{h^{0.5}}\int_0^1 K([w-s]/h)dU_N^H(s)$$
$$= \sqrt{Nh}\left\{\frac{1}{h}\int_0^1 K([w-s]/h)d\Gamma_N^H(s) - \frac{1}{h}\int_0^1 K([w-s]/h)ds\right\}.$$

The last term is equal to 1 for h small enough, so one has

$$A_{2N} = \sqrt{Nh}\Big\{\frac{1}{Nh}\sum_{i=1}^{m} K([w-U_i]/h) + \frac{1}{Nh}\sum_{j=1}^{n} K([w-V_j]/h) - 1\Big\},$$

where $V_j = F(Y_j)$. Now examine $A_{1N} - A_{2N}$:

$$\begin{aligned}
&A_{1N} - A_{2N} \\
&= \sqrt{Nh}\Big\{\frac{1}{mh}\sum_{i=1}^{m} K([w-U_i]/h) - \frac{1}{Nh}\sum_{i=1}^{m} K([w-U_i]/h) \\
&\quad - \frac{1}{Nh}\sum_{j=1}^{n} K([w-V_j]/h) - 1 + 1\Big\} \\
&= \sqrt{Nh}\Big\{\frac{1-\lambda_{(N)}}{mh}\sum_{i=1}^{m} K([w-U_i]/h) - \frac{1-\lambda_{(N)}}{nh}\sum_{j=1}^{n} K([w-V_j]/h)\Big\} \\
&= \frac{1-\lambda_{(N)}}{\sqrt{\lambda_{(N)}}}\sqrt{mh}\Big\{\frac{1}{mh}\sum_{i=1}^{m} K([w-U_i]/h) - 1\Big\} \\
&\quad - \sqrt{\frac{nh}{1-\lambda_{(N)}}}\Big\{\frac{1}{nh}\sum_{j=1}^{n} K([w-V_j]/h) - 1\Big\}.
\end{aligned}$$

Both terms in this last equality can be shown to converge to a limiting normal distribution by invoking a CLT for triangular arrays. Since both terms are independent, one can find the limiting distribution of the sum by finding the sum of the limiting distributions. One reaches the conclusion that

$$A_{1N} - A_{2N} \xrightarrow{\mathrm{d}} Z$$

where Z has the normal distribution with mean 0 and variance

$$\frac{1-\lambda_0}{\lambda_0}\int_{-1}^{1} K(t)^2 dt$$

since $h \to 0$, $(m \wedge n)h \to \infty$ as $m \wedge n \to \infty$.

Next it is shown that B_{1N} converges to 0 in probability. Start by bounding B_{1N}:

$$|B_{1N}| \le \sup_{0\le s\le 1} |U_N^H(s)| \frac{1}{h^{1.5}} \int_0^1 \Big|K'([w-t_N(s)]/h) - K'([w-s]/h)\Big| ds.$$

For $\Gamma_N^H(s)$ and s outside the range $(w-h, w+h)$, the integrand is identically zero. Inside this range, the mean value theorem and the assumption that the second derivative of K is bounded by M give

$$\left|K'([w-t_N(s)]/h) - K'([w-s]/h)\right| \leq \frac{M}{h}|t_N(s)-s|.$$

This yields the following bound on B_{1N}:

$$|B_{1N}| \leq M \sup_{0\leq s\leq 1}|U_N^H(s)|\frac{1}{h^{1.5}}\int_0^1 \frac{1}{h}|t_N(s)-s|I_h(s,\Gamma_N^H(s))ds$$

$$\text{(B.2)} \qquad \leq M(\sup_{0\leq s\leq 1}|U_N^H(s)|)^2\frac{1}{\sqrt{Nh^3}}\int_0^1 \frac{1}{h}I_h(s,\Gamma_N^H(s))ds,$$

since $|t_N(s)-s| \leq |\Gamma_N^H(s)-s|$ and where $I_h(s,\Gamma_N^H(s))$ is 1 if either $w-h \leq s \leq w+h$ or $w-h \leq \Gamma_N^H(s) \leq w+h$. An equivalent expression for $I_h(s,\Gamma_N^H(s))$ is

$$\text{(B.3)} \qquad I_h(s,\Gamma_N^H(s)) = I_h(s) + I_h(\Gamma_N^H(s)) - I_h(s)I_h(\Gamma_N^H(s)),$$

where $I_h(r) = 1$ if $w-h \leq r \leq w+h$ and 0 otherwise.

The last integral in (B.2) must be evaluated. Substituting the expression (B.3) into this integral results in

$$\int_0^1 \frac{1}{h}I_h(s,\Gamma_N^H(s))ds = 2 + \int_0^1 \frac{1}{h}I_h(\Gamma_N^H(s))ds - \int_{w-h}^{w+h}\frac{1}{h}I_h(\Gamma_N^H(s))ds$$
$$\equiv 2 + R_{1N} + R_{2N}.$$

The function $\Gamma_N^H(s)$ is the empirical distribution function of $F(X_1),\ldots,F(X_m)$, and $F(Y_1),\ldots,F(Y_n)$ which are distributed as N iid U(0,1) random variables under H_o. One can bound R_{1N} by $[Q_N(w+1.5h) - Q_N(w-1.5h)]/h$, where Q_N is the empirical quantile function of these N random variables. This last bound looks like a derivative. Applying Bahadhur's representation to the sample quantiles, one can show that $[Q_N(w+1.5h) - Q_N(w-1.5h)]/h \xrightarrow{p} 3$ since $(m \wedge n)h^2 \to \infty$. The term R_{2N} is easily handled since $0 \leq R_{2N} \leq 2$.

Putting all the components together, one concludes that

$$|B_{1N}| \leq M(\sup_{0\leq s\leq 1}|U_N^H(s)|)^2\frac{1}{\sqrt{Nh^3}}\left[\int_0^1 \frac{1}{h}I_h(s,\Gamma_N^H(s))ds\right]$$

$$= O_p(1)/\sqrt{Nh^3} \xrightarrow{p} 0$$

since $(m \wedge n)h^3 \to \infty$. The term $\sup_{0\le s\le 1}|U_N^H(s)|$ is $O_p(1)$ since it converges in distribution as $m \wedge n \to \infty$. In fact, this term is the Kolmogorov-Smirnov statistic; see Shorack and Wellner (1986), page 91.

It must now be shown that $B_{2N} \xrightarrow{p} 0$. The term B_{2N} is written as $|B_{2N}| \le |C_{1N}| + |C_{2N}|$ where

$$C_{1N} = \left|\frac{1}{h^{1.5}}\int_0^1 \left[K'([w - t_N(s)]/h) - K'([w-s]/h)\right] U_N^H(s)d[\Gamma_m^F(s) - s]\right|,$$

and

$$C_{2N} = \left|\frac{1}{h^{1.5}}\int_0^1 K'([w-s]/h)U_N^H(s)d[\Gamma_m^F(s) - s]\right|.$$

Start by examining C_{1N}. It can be shown that C_{1N} is bounded by

$$\begin{aligned} C_{1N} &\le \sup_{0\le s\le 1}|U_N^H(s)|\frac{1}{h^{1.5}}\int_0^1 \frac{M}{h}|t_N(s) - s|I_h(s,\Gamma_s^H(s))d\Gamma_M^F(s) \\ &\quad + \sup_{0\le s\le 1}|U_N^H(s)|\frac{1}{h^{1.5}}\int_0^1 \frac{M}{h}|t_N(s) - s|I_h(s,\Gamma_N^H(s))ds \\ &\le \left(\sup_{0\le s\le 1}|U_N^H(s)|\right)^2\frac{M}{\sqrt{Nh^3}}\left\{\int_0^1 \frac{1}{h}I_h(s,\Gamma_N^H(s))d\Gamma_m^F(s)\right. \\ &\quad \left. + \int_0^1 \frac{1}{h}I_h(s,\Gamma_N^H(s))ds\right\}. \end{aligned} \tag{B.4}$$

From above it is known that the second term in the brackets is $O_p(1)$. Examine the first term in the brackets:

$$\begin{aligned} &\int_0^1 \frac{1}{h}I_h(s,\Gamma_N^H(s))d\Gamma_m^F(s) \\ &= \int_0^1 \frac{1}{h}I_h(s)d\Gamma_m^F(s) + \frac{1}{h}\int_0^1 I_h(\Gamma_N^H(s))d\Gamma_m^F(s) \\ &\quad - \int_0^1 \frac{1}{h}I_h(s)I_h(\Gamma_N^H(s))d\Gamma_m^F(s) \end{aligned} \tag{B.5}$$

The first term of (B.5) is equal to

$$\frac{1}{mh}\sum_{i=1}^m I_h(U_i)$$

which is a kernel density estimate at the point w using the uniform kernel. It converges in probability [cf. Parzen (1962a)] to 2. The second term of (B.5) is

handled by

$$
\begin{aligned}
0 &\leq \int_0^1 \frac{1}{h} I_h(\Gamma_N^H(s)) d\Gamma_m^F(s) \\
&= \frac{1}{mh} \sum_{i=1}^{m} I_h(\Gamma_N^H(U_i)) \\
&\leq \frac{1}{mh} \sum_{i=1}^{N} I_h(\Gamma_N^H(Z_i)) \\
&= \frac{1}{mh} \sum_{i=1}^{N} I_h(i/N) \to 2/\lambda_0,
\end{aligned}
$$

where $Z_i = U_i$ for $i = 1, \ldots, m$ and $Z_{i+m} = V_i$ for $i = 1, \ldots, n$. Hence the second term is $O_p(1)$. Similarly, for the third term of (B.5) one has

$$
\begin{aligned}
0 &\leq \int_0^1 \frac{1}{h} I_h(s) I_h(\Gamma_N^H(s)) d\Gamma_m^F(s) \\
&\leq \int_0^1 \frac{1}{h} I_h(s) d\Gamma_m^F(s) = O_p(1)
\end{aligned}
$$

Returning to equation (B.4), the term in braces has been shown to be $O_p(1)$. Since $(m \wedge n)h^3 \to \infty$ then $C_{1N} \xrightarrow{\text{p}} 0$ as $m \wedge n \to \infty$.

Now examine C_{2N}. This is the more difficult of the two terms. After some rearranging, C_{2N} can be written as

$$
\begin{aligned}
C_{2N} = -\sqrt{Nh}\Big[&\frac{1}{Nh} \sum_{i=1}^{N} K([w - Z_i]/h) - 1 \\
&- \frac{1}{mh^2} \sum_{i=1}^{m} K'([w - U_i]/h)[\Gamma_N^H(U_i) - U_i]\Big].
\end{aligned}
$$

By repeated application of the mean value theorem, one finds that

$$
\begin{aligned}
D_{1N} = \sqrt{Nh}\Big[&\frac{1}{Nh} \sum_{i=1}^{N} K([w - Z_i]/h) - \frac{1}{Nh} \sum_{i=1}^{N} K([w - (i/N)]/h) \\
&- \frac{1}{Nh^2} \sum_{i=1}^{N} K'([w - r_{iN}]/h)[\Gamma_N^H(Z_i) - Z_i]\Big] = 0,
\end{aligned}
$$

where r_{iN} is between Z_i and $\Gamma_N^H(Z_i)$ and since $\Gamma_N^H(Z_{(i)}) = i/N$. If it can be shown that

$$D_{2N} = \sqrt{Nh}\Big[\frac{1}{Nh}\sum_{i=1}^{N} K([w-(i/N)]/h) - 1\Big] \to 0,$$

$$D_{3N} = \sqrt{Nh}\Big[\frac{1}{Nh^2}\sum_{i=1}^{N} K'([w-r_{iN}]/h)[\Gamma_N^H(Z_i) - Z_i]$$

$$- \frac{1}{Nh^2}\sum_{i=1}^{N} K'([w-Z_i]/h)[\Gamma_N^H(Z_i) - Z_i]\Big] \xrightarrow{\mathrm{P}} 0,$$

$$D_{4N} = \sqrt{Nh}\Big[\frac{1}{Nh^2}\sum_{i=1}^{N} K'([w-Z_i]/h)[\Gamma_N^H(Z_i) - Z_i]$$

$$- \frac{1}{mh^2}\sum_{i=1}^{m} K'([w-U_i]/h)[\Gamma_N^H(U_i) - U_i]\Big] \xrightarrow{\mathrm{P}} 0,$$

then $C_{2N} \xrightarrow{\mathrm{P}} 0$ since $C_{2N} = -D_{1N} - D_{2N} - D_{3N} - D_{4N}$. Start with D_{2N}. Let

$$g(t) = \int_0^t \frac{1}{h} K([w-s]/h)ds.$$

By Taylor's theorem $g(t) = g(r) + g'(r)(t-r) + 0.5(t-r)^2 g''(c)$ where c is between t and r. Let $t = (i-1)/N$ and $r = i/N$. Then

$$g([i-1]/N) = g(i/N) - \frac{1}{Nh}K([w-(i/N)]/h) + O(1/N^2h^2),$$

if $i/N > w - h$ and $(i-1)/N < w + h$; otherwise $g(i/N) = g([i-1]/N) = 0$. So

$$\sum_{i=1}^{N}[g(i/N) - g([i-1]/N) - \frac{1}{Nh}K([w-(i/N)]/h)$$

$$= \sum_{i=1}^{N} O(1/N^2h^2)I(i/N > w-h)I([i-1]/N < w+h).$$

There are about $2Nh$ terms in this last sum, so $2NhO(1/N^2h^2) = O(1/Nh)$ and

$$\sum_{i=1}^{N}[g(i/N) - g([i-1]/N) - \frac{1}{Nh}K([w-(i/N)]/h)$$

$$= g(1) - g(0) - \frac{1}{Nh}\sum_{i=1}^{N} K([w-(i/N)]/h)$$

$$= 1 - \frac{1}{Nh}\sum_{i=1}^{N} K([w-(i/N)]/h) = O(1/Nh).$$

This implies that $D_{2N} = O(1/\sqrt{Nh})$ and hence $D_{2N} \to 0$ since $(m \wedge n)h \to \infty$.

Next examine D_{3N}. This term can be bounded by

$$|D_{3N}| \leq \sqrt{Nh} \sup_{0\leq s\leq 1} |\Gamma_N^H(s) - s|$$
$$\cdot \frac{1}{Nh^2}\sum_{i=1}^{N} |K'([w-r_{iN}]/h) - K'([w-Z_i]/h)|$$

$$\leq \sup_{0\leq s\leq 1} |U_N^H(s)| \frac{M}{Nh^{1.5}} \sum_{i=1}^{N} \frac{1}{h}|r_{iN} - Z_i|$$
$$\cdot \Big\{ I_h(Z_i) + I_h(\Gamma_N^H(Z_i)) - I_h(Z_i) I_h(\Gamma_N^H(Z_i)) \Big\}.$$

The last sum results from the mean value theorem and the fact that $K'([w - r_{iN}]/h) = 0$ and $K'([w - Z_i]/h) = 0$ if both Z_i and $\Gamma_N^H(Z_i)$ are outside the interval $(w-h, w+h)$. The first term in the braces is the kernel density estimate of the uniform density (under H_o) based on the Z_i's. This quantity converges in probability to 2 as $(m \wedge n)h \to \infty$. The second term is actually nonrandom and is just

$$\frac{1}{Nh}\sum_{i=1}^{N} I_h(i/N)$$

which converges to 2 as $(m \wedge n)h \to \infty$. The last term in the braces is bounded by both the first and second terms. Thus

$$|D_{3N}| \leq \Big(\sup_{0\leq s\leq 1} |U_N^H(s)| \Big)^2 \frac{1}{\sqrt{Nh^3}} O_p(1)$$

and so one may conclude that $D_{3N} \xrightarrow{p} 0$ since $(m \wedge n)h^3 \to \infty$.

The term D_{4N} is the most tedious to deal with. It can be rewritten as

$$D_{4N} = (1 - \lambda_{(N)}) \Big[\frac{1}{mh^{1.5}} \sum_{i=1}^{m} K'([w-U_i]/h) U_N^H(U_i)$$
$$- \frac{1}{nh^{1.5}} \sum_{i=1}^{n} K'([w-V_i]/h) U_N^H(V_i) \Big]$$

The claim is that $D_{4N} \xrightarrow{\text{ms}} 0$. One needs to show that $\mathrm{E}[D_{4N}^2] \to 0$. Begin by evaluating the expectation

$$\mathrm{E}\Big[\frac{1}{h^3}K'([w-Z_i]/h)K'([w-Z_j]/h)U_N^H(Z_i)U_N^H(Z_j)\Big].$$

Start for $i \neq j$ and first find the expectation

$$\mathrm{E}[U_N^H(Z_1)U_N^H(Z_2)|Z_1 = s, Z_2 = t]$$

After several pages of algebra, this expectation is found to be

$$\min(s,t) - st + O(1/N).$$

One can then find the desired expectation from the conditional expectation:

$$\begin{aligned}
&\mathrm{E}\Big[\frac{1}{h^3}K'([w-Z_1]/h)K'([w-Z_2]/h)U_N^H(Z_1)U_N^H(Z_2)\Big] \\
&= \mathrm{E}_{Z_1,Z_2}\Big[\frac{1}{h^3}K'([w-Z_1]/h)K'([w-Z_2]/h)\mathrm{E}\Big[U_N^H(Z_1)U_N^H(Z_2)|Z_1,Z_2\Big]\Big] \\
&= \frac{1}{h}\int_{-1}^{1}\int_{-1}^{1}K'(u)K'(v)\Big[\min(w-hu,w-hv) - (w-hu)(w-hv)\Big]dudv \\
&\quad + O(1/Nh) \\
&\equiv C(h),
\end{aligned}$$

for $i \neq j$. The procedure is similar for $i = j$ and one arrives at

$$\begin{aligned}
&\mathrm{E}\Big[\frac{1}{h^3}K'([w-Z_1]/h)^2U_N^H(Z_1)^2\Big] \\
&= \frac{1}{h^2}\int_{-1}^{1}K'(y)^2(w-hy)(1-w+hy)dy + O(1/Nh^2) \\
&\equiv V(h)
\end{aligned}$$

Putting these results together, one finds

$$\mathrm{E}[D_{4N}^2] = \frac{V(h)}{m} + \frac{V(h)}{n} - \frac{C(h)}{m} - \frac{C(h)}{n}.$$

It is obvious that

$$\begin{aligned}
&\frac{V(h)}{n} \to 0 \text{ if } nh^2 \to \infty, \\
&\frac{V(h)}{m} \to 0 \text{ if } mh^2 \to \infty, \\
&\frac{C(h)}{n} \to 0 \text{ if } nh \to \infty, \\
&\frac{C(h)}{n} \to 0 \text{ if } mh \to \infty,
\end{aligned}$$

and hence that $D_{4N} \xrightarrow{\text{ms}} 0$ which implies that $D_{4N} \xrightarrow{\text{p}} 0$. Backing up through the remainder terms, it has been shown now that $C_{2N} \xrightarrow{\text{p}} 0$ so that $B_{2N} \xrightarrow{\text{p}} 0$. An application of Slutsky's theorem to A_{1N}, A_{2N}, B_{1N}, and B_{2N} yields the desired result.

Proof of Theorem 3.2.2

Theorem 3.2.2 is concerned with the consistency of the boundary kernel estimator of the comparison density function. Start by writing

$$|\hat{d}_h(w) - d_o(w)| \leq A_{1N} + A_{2N} + A_{3N},$$

where

$$A_{1N} = \frac{1}{\sqrt{N}}\left|\frac{1}{h}\int_0^1 K([w-u]/h)d\mathrm{CD}_N(u)\right|,$$

$$A_{2N} = \left|\frac{1}{h}\int_0^1 K([w-u]/h)dD_{(N)}(u) - d_{(N)}(w)\right|,$$

$$A_{3N} = |d_{(N)}(w) - d_o(w)|.$$

As in Theorem 3.2.1, since $h \to 0$ it is not necessary to worry about boundary modifications. Each of these terms must be shown to converge to zero in probability. Start with A_{1N}:

$$A_{1N} \leq \frac{1}{\sqrt{N}h^2}\int_0^1 |\mathrm{CD}_N(u)||K'([w-u]/h)|du$$

$$\leq \sup_{0\leq s\leq 1}|\mathrm{CD}_N(u)|\frac{1}{\sqrt{N}h}\int_{-1}^1 |K'(y)|dy.$$

One can show $\sup_{0\leq s\leq 1}|\mathrm{CD}_N(u)|$ is $O_p(1)$ in the following manner. It is known (see Section 2) that CD_N converges to a limiting process under fixed alternatives as well as under H_o. By Theorem 3.11 of Ruymgaart (1988) it follows that this term converges in distribution to a limiting random variable and hence it is $O_p(1)$. It follows that $A_{1N} \xrightarrow{\text{p}} 0$ since $(m \wedge n)h^2 \to \infty$ by assumption.

The proof that $A_{2N} \to 0$ follows that of Theorem 1A of Parzen (1962a). Let

$$g_{(N)}(w) = \frac{1}{h}\int_0^1 K([w-u]/h)d_{(N)}(u)du$$

$$= \int_{-\infty}^{\infty}\frac{1}{h}K(y/h)d_{(N)}(w-y)dy$$

for h sufficiently small. Pick $\delta > 0$ and write

$$
\begin{aligned}
&|g_{(N)}(w) - d_{(N)}(w)| \\
&\leq \Big| \int_{|y|\leq\delta} |d_{(N)}(w-y) - d_{(N)}(w)| \frac{1}{h} K(y/h) dy \Big| \\
&\quad + \Big| \int_{|y|>\delta} |d_{(N)}(w-y) - d_{(N)}(w)| \frac{1}{h} K(y/h) dy \Big| \\
&\leq \max_{|y|\leq\delta} |d_{(N)}(w-y) - d_{(N)}(w)| \int_{|z|\leq\delta/h} |K(z)| dz \\
&\quad + \int_{|y|>\delta} \frac{1}{y} d_{(N)}(w-y) \frac{y}{h} K(y/h) dy \\
&\quad + d_{(N)}(w) \int_{|y|>\delta} \frac{1}{h} K(y/h) dy \\
&\leq \max_{|y|\leq\delta} |d_{(N)}(w-y) - d_{(N)}(w)| \int_{|z|<\delta/h} \\
&\quad + \frac{1}{\delta} \sup_{|z|>\delta/h} |zK(z)| \int_{-\infty}^{\infty} d_{(N)}(y) dy + d_{(N)}(w) \int_{|z|>\delta/h} K(z) dz.
\end{aligned}
$$

The strategy is to let $m \wedge n \to \infty$ for fixed δ and then let $\delta \to 0$. The last two terms tend to zero as $m \wedge n \to \infty$ and $h \to 0$. This leaves only the first term. Rewrite the first term as follows:

$$
\begin{aligned}
\max_{|y|\leq\delta} |d_{(N)}(w-y) - d_{(N)}(w)| &\leq \max_{|y|\leq\delta} |d_{(N)}(w-y) - d_o(w-y)| \\
&\quad + \max_{|y|\leq\delta} |d_o(w) - d_{(N)}(w)| + \max_{|y|\leq\delta} |d_o(w-y) - d_o(w)|.
\end{aligned}
$$

The second term tends to zero as $m \wedge n \to \infty$ because $d_{(N)}$ is converging to d_o. The third term tends to zero as $\delta \to 0$ because d_o is continuous. This leaves only the first term. It will tend to zero if $d_{(N)}$ converges to d_o uniformly. If $\lambda_{(N)} = \lambda_0$, then $d_{(N)} = d_o$ and Theorem 1A of Parzen (1962a) can be applied directly to show that $A_{2N} \to 0$. Since $d_{(N)}$ converges to d_o, $A_{3N} \to 0$.

Conclude that under these conditions, one has $\hat{d}_h(w) \xrightarrow{p} d_o(w)$.

Proof of Lemma 3.2.1

Lemma 3.2.1 states that the Gasser-Müller boundary kernel satisfies the Regularity Conditions. It must be shown that for all $\epsilon > 0$ there exists a $\delta = \delta(\epsilon)$

such that

$$\int_0^1 |K_s'([w-u]/h) - K_{s'}'([v-u]/h)|du < \epsilon$$

if $|w-v| < \delta$. Define

$$P = \{u : A(w,[w-u]/h) = A(v,[v-u]/h) = 1\},$$
$$Q = \{u : A(w,[w-u]/h) \neq A(v,[v-u]/h)\},$$

where

$$A(s,t) = \begin{cases} I[-1,1](t) & \text{if } h \leq s \leq 1-h \\ I[(s-1)/h,1](t) & \text{if } 1-h \leq s \leq 1 \\ I[-1,s/h](t) & \text{if } 0 \leq s \leq h, \end{cases}$$

and I is an indicator function. The function $A(s,t)$ is indicating the region where the boundary kernel is non-zero. One can now write

$$\begin{aligned} &\int_0^1 |K_s'([w-u]/h) - K_{s'}'([v-u]/h)|du \\ &= \int_P |K_s'([w-u]/h) - K_{s'}'([v-u]/h)|du \\ &\quad + \int_Q |K_s'([w-u]/h) - K_{s'}'([v-u]/h)|du \\ &\leq \sup_{u\in P} |K_s'([w-u]/h) - K_{s'}'([v-u]/h)| \\ &\quad + \int_Q |K_s'([w-u]/h) - K_{s'}'([v-u]/h)|du, \end{aligned} \tag{B.6}$$

where $s = s(w,h)$ and $s' = s(v,h)$ index the boundary kernel (see Section 2). Concentrate on the first term of (B.6). For $u \in P$, one has $A(w,[w-u]/h) = 1$ and $A(v,[v-u]/h) = 1$, so that

$$\begin{aligned} &\sup_{u\in P} |K_s'([w-u]/h) - K_{s'}'([v-u]/h)| \\ &= \sup_{u\in P} |(\theta_s + \phi_s(w-u)/h)K'([w-u]/h) + \phi_s K([w-u]/h) \\ &\quad - (\theta_{s'} + \phi_{s'}(v-u)/h)K'([v-u]/h) - \phi_{s'}K([v-u]/h)| \\ &\leq (M_1 + M_2 + M_3 + 2M_4 + M_5)\epsilon^*, \end{aligned}$$

if

$$|w-v| < h\delta(\epsilon^*) = h\min(\delta_1(\epsilon^*), \delta_2(\epsilon^*), \delta_3(\epsilon^*), \delta_4(\epsilon^*), \delta_5(\epsilon^*)),$$

where

$$M_1 = \sup_u |K'(u)|$$
$$M_2 = \sup_{0 \le s \le 1} |\theta_s|$$
$$M_3 = \sup |sK'(s)|$$
$$M_4 = \sup_{0 \le s \le 1} |\phi_s|$$
$$M_5 = \sup |K(s)|.$$

The $\delta(\epsilon^*)$'s result from the uniform continuity (u.c.) of the following functions:

$$\delta_1(\epsilon^*) : \theta_s \text{ is u.c. in } s$$
$$\delta_2(\epsilon^*) : K'(t) \text{ is u.c.}$$
$$\delta_3(\epsilon^*) : \phi_s \text{ is u.c. in } s$$
$$\delta_4(\epsilon^*) : tK'(t) \text{ is u.c.}$$
$$\delta_5(\epsilon^*) : K \text{ is u.c.}$$

Thus one can choose ϵ^* so that

$$\text{(B.7)} \qquad \sup_{s \in P} |K'_s([w-u]/h) - K'_{s'}([v-u]/h)| < \epsilon/2$$

if $|w - v| < \delta(\epsilon)$.

Moving to the second term of equation (B.6), the claim is that Q is either (1) empty, (2) an interval of length less than $|w - v|$, or (3) the sum of two intervals the sum of whose length is less than $2|w - v|$.

Define $|Q|$ to be the Lebesgue measure of Q. The above claim is proved by enumerating all possible combinations of cases of the boundary kernel: left boundary, interior, and right boundary.

1. If $w = v$ then Q is empty.
2. If $0 \le w \le h$ then $K'_s([w-u]/h)$ is nonzero on $0 \le u \le w + h$; if $0 \le v \le h$ then $K'_{s'}([v-u]/h)$ is nonzero on $0 \le u \le v + h$. Thus Q has measure $|w - v|$.
3. If $1 - h \le w \le 1$ then $K'_s([w-u]/h)$ is nonzero on $w - h \le u \le 1$; if $1 - h \le v \le 1$ then $K'_{s'}([v-u]/h)$ is nonzero on $v - h \le u \le 1$. Then $|Q| = |w - v|$.

4. If $0 \le w \le h$ then $K_s'([w-u]/h)$ is nonzero on $0 \le u \le w+h$; if $h \le v \le 1-h$ then $K_{s'}'([v-u]/h)$ is nonzero on $v-h \le u \le v+h$. If the two supports don't overlap, then $|Q| = 3h + w \le 2|w-v|$. If the two supports do overlap, then

$$|Q| = |w-v| + |v-h| \le 2|w-v|.$$

5. If $h \le w \le 1-h$ then $K_s'([w-u]/h)$ is nonzero on $w-h \le u \le w+h$; if $h \le v \le 1-h$ then $K_{s'}'([v-u]/h)$ is nonzero on $v-h \le u \le v+h$. If the two supports don't overlap, then $|Q| = 4h \le 2|w-v|$. If the two supports do overlap, then $|Q| = 2|w-v|$.
6. If $1-h \le w \le 1$ then $K_s'([w-u]/h)$ is nonzero on $w-h \le u \le 1$; if $h \le v \le 1-h$ then $K_{s'}'([v-u]/h)$ is nonzero on $v-h \le u \le v+h$. If the two supports don't overlap, then $|Q| = 3h + 1 - w \le 2|w-v|$. If the two supports do overlap, then

$$|Q| = |w-v| + |1-h-v| \le 2|w-v|.$$

7. If $0 \le w \le h$ then $K_s'([w-u]/h)$ is nonzero on $0 \le u \le w+h$; if $1-h \le v \le 1$ then $K_{s'}'([v-u]/h)$ is nonzero on $v-h \le u \le 1$. If the two supports don't overlap, then $|Q| = 1 + 2h + w - v \le 4h \le 2|w-v|$. If the two supports do overlap, then

$$|Q| = |v-h| + |1-h-w| \le 2|w-v|.$$

Hence $|Q| \le 2|w-v|$. This implies that

$$\int_Q |K_s'([w-u]/h) - K_{s'}'([v-u]/h)du \le 4M|w-v| \tag{B.8}$$

where $M = \sup_{s,t} |K_s'(t)| < \infty$ because K' is continuous and $\delta(\epsilon) = \epsilon/8M$.

Finally, choose

$$\delta^*(\epsilon) = \min(\delta(\epsilon), \epsilon/8M).$$

Combining (B.7) and (B.8) with this choice of $\delta(\epsilon)$ one sees that

$$\int_0^1 |K_s'([w-u]/h) - K_{s'}'([v-u]/h)|du < \epsilon$$

if $|w - v| < \delta^*(\epsilon)$. Hence, the result is uniform continuity and the Regularity Condition holds.

Proof of Lemma 3.2.2

It is to be proved that the sample paths of KDP_h exist and are continuous with probability 1. Start the proof by showing that $L(u)$ is continuous with probability 1. Pyke and Shorack (1968) give the following representation for $L(u)$:

$$L(u) = (1 - \lambda_0)\Big\{\frac{1}{\sqrt{\lambda_0}} d_o^G(u) U_o[D_o^F(u)] - \frac{1}{\sqrt{1-\lambda_0}} d_o^F(u) V_o[D_o^G(u)]\Big\},$$

where U_o and V_o are independent Brownian bridges and

$$\begin{aligned} D_o^F(u) &= FQ_o^H(u), \\ D_o^G(u) &= GQ_o^H(u), \\ d_o^F(u) &= \frac{d}{du} D_o^F(u), \\ d_o^G(u) &= \frac{d}{du} D_o^G(u), \\ Q_o^H(u) &= H_o^{-1}(u), \\ H_o(x) &= \lambda_0 F(x) + (1 - \lambda_0) G(x). \end{aligned}$$

It is known [see Billingsley (1968), page 61] that the sample paths of a Brownian bridge are continuous with probability 1. As a result of the assumptions made about the two distributions, the functions D^F, D^G, d^F, and d^G are continuous. Since the compositions, products and differences of continuous functions are also continuous, conclude that the sample paths of $L(u)$ are continuous with probability 1.

Since $L(u)$ is continuous with probability 1, the integral defining KDP_h exists with probability 1. Define $r(w)$ by

$$r(w) = \frac{1}{h^2} \int_0^1 K'([w - u]/h) c(u) du,$$

where $c(u)$ is any continuous function. It must be shown that $r(w)$ is continuous. This is done by bounding $|g(w) - g(v)|$ as follows:

$$|g(w) - g(v)| \leq \sup_{0 \leq u \leq 1} |c(u)| \int_0^1 |K'([w - u]/h) - K'([v - u]/h)| du.$$

Since $c(u)$ is continuous, $\sup_{0\leq u\leq 1}|c(u)| < \infty$. It was shown in Lemma 3.2.1 that the integral can be made smaller than any $\epsilon > 0$ if $|w - v| < \delta(\epsilon)$. Hence $g(w)$ is continuous for any continuous function $c(u)$. Thus, it may be concluded that KDP_h is continuous with probability 1.

Proof of Theorem 3.2.3

The proof of Theorem 3.2.3 is patterned after the proof of Theorem 3.9 of Ruymgaart (1988). This latter theorem concerns the weak convergence of the uniform empirical process to a Brownian bridge process. Before proving Theorem 3.2.3, a lemma is stated and proved.

Lemma B.1. *Let*

$$g(w) = \frac{1}{h^2}\int_0^1 U(t)K_s'([w-t]/h)dt,$$

where $U \in D[0,1]$; $K_s'(t)$ is the first derivative of a boundary kernel satisfying the regularity conditions and $s = s(w,h)$. Then $g(w)$ is uniformly continuous and

$$\sup_{|w-v|<\delta}|g(w) - g(v)| \leq \frac{1}{h^2}\sup_{0\leq t\leq 1}|U(t)|\theta(\delta),$$

where $\theta(\delta)$ is defined in the statement of the Regularity Conditions in Subsection 3.2.2.

Proof.

The term $|g(w) - g(v)|$ can be bounded by

$$\frac{1}{h^2}\sup_{0\leq t\leq 1}|U(t)|\int_0^1 |K_s([w-t]/h) - K_{s'}([v-t]/h)|dt.$$

Taking the supremum of the above over $|w - v| < \delta$ yields

$$\sup_{|w-v|<\delta}|g(w) - g(v)| \leq \frac{1}{h^2}\sup_{0\leq t\leq 1}|U(t)|\theta(\delta).$$

From this it follows that $g(w)$ is uniformly continuous since the bound on the right does not depend on w or v. The bound on the right is finite since $U \in D[0,1]$ implies that $\sup_{0\leq t\leq 1}|U(t)| < \infty$ [see Gaenssler (1983), page 90, for a statement of this].

The strategy for proving the weak convergence of $\mathrm{KDP}_{N,h}$ to KDP_h is to show that

$$\mathrm{E}[g(\mathrm{KDP}_{N,h})] \to \mathrm{E}[g(\mathrm{KDP}_h)] \text{ as } m \wedge n \to \infty,$$

where g is any bounded and ρ-uniformly continuous functional, $g: C[0,1] \to \mathbb{R}$. The norm ρ is taken to be the sup-norm.

Start the proof of Theorem 3.2.3 by defining C_ρ^* to be

$$C_\rho^* = \Big\{ g: C[0,1] \to \mathbb{R} : g \text{ is } \rho-\text{uniformly continuous, bounded and measurable.} \Big\},$$

and choose g to be any function in C_ρ^*. Define

$$\delta_g(\epsilon) = \sup_{\rho(s,t)<\epsilon} |g(t) - g(s)|,$$

$$C = \sup_t |g(t)|,$$

$$A_l(g;s) = \begin{cases} g(s) & \text{for } s = 0/l, 1/l, \ldots, (l-1)/l, l/l, \\ l\Big[(s - [i-1]/l)g(i/l) & \\ \quad +(i/l - s)g([i-1]/l)\Big] & \text{for } [i-1]/l < s < i/l. \end{cases}$$

It is easily seen that $A_l(g;s)$ is piece-wise continuous with nodes at $(i/l, g(i/l))$.

To show that the expected value converges, the quantity

$$\Big|\mathrm{E}[g(\mathrm{KDP}_{N,h})] - \mathrm{E}[g(\mathrm{KDP}_h)]\Big|$$

will be shown to be bounded by terms decreasing to zero. Start by applying the triangle inequality:

$$\begin{aligned} &\Big|\mathrm{E}[g(\mathrm{KDP}_{N,h})] - \mathrm{E}[g(\mathrm{KDP}_h)]\Big| \\ &\le \Big|\mathrm{E}[g(\mathrm{KDP}_{N,h})] - \mathrm{E}[g(A_l(\mathrm{KDP}_{N,h}))]\Big| \\ &\quad + \Big|\mathrm{E}[g(A_l(\mathrm{KDP}_{N,h}))] - \mathrm{E}[g(A_l(\mathrm{KDP}_h))]\Big| \\ &\quad + \Big|\mathrm{E}[g(A_l(\mathrm{KDP}_h))] - \mathrm{E}[g(\mathrm{KDP}_h)]\Big| \\ &\le \delta_g(\epsilon) + 2C\mathrm{P}[\rho(\mathrm{KDP}_{N,h}, A_l(\mathrm{KDP}_{N,h})) > \epsilon] \\ &\quad + \Big|\mathrm{E}[g(A_l(\mathrm{KDP}_{N,h}))] - \mathrm{E}[g(A_l(\mathrm{KDP}_h))]\Big| \\ &\quad + \delta_g(\epsilon) + 2C\mathrm{P}[\rho(\mathrm{KDP}_h, A_l(\mathrm{KDP}_h)) > \epsilon]. \end{aligned} \tag{B.9}$$

The first two terms of (B.9) result by splitting the space S_D^h into two subspaces: one where

$$\rho(\mathrm{KDP}_{N,h}, A_l(\mathrm{KDP}_{N,h})) < \epsilon$$

and one where

$$\rho(\mathrm{KDP}_{N,h}, A_l(\mathrm{KDP}_{N,h})) \geq \epsilon.$$

On the first region, $|g(\mathrm{KDP}_{N,h}) - g(A_l(\mathrm{KDP}_{N,h}))| < \delta_g(\epsilon)$ and so the expected values differ by less than $\delta_g(\epsilon)$. On the second region,

$$|g(\mathrm{KDP}_{N,h}) - g(A_l(\mathrm{KDP}_{N,h}))| < 2C,$$

and so the second term results. The last two terms of (B.9) are derived in an identical fashion.

First handle the term $\mathrm{P}[\rho(\mathrm{KDP}_{N,h}, A_l(\mathrm{KDP}_{N,h})) > \epsilon]$. By Lemma B.1, one has

$$\sup_{|w-v|<\delta} |\mathrm{KDP}_{N,h}(w) - \mathrm{KDP}_{N,h}(v)| \leq \frac{1}{h^2} \sup_{0\leq t\leq 1} |\mathrm{CD}_N(t)|\theta(\delta).$$

By the construction of A_l, it can be seen that

$$\rho(\mathrm{KDP}_{N,h}, A_l(\mathrm{KDP}_{N,h})) \leq \sup_{|w-v|<1/l} \left|\mathrm{KDP}_{N,h}(w) - \mathrm{KDP}_{N,h}(v)\right|.$$

These two facts lead to:

$$\begin{aligned}
&\mathrm{P}[\rho(\mathrm{KDP}_{N,h}, A_l(\mathrm{KDP}_{N,h})) > \epsilon] \\
&\leq \mathrm{P}[\frac{1}{h^2} \sup_{0\leq t\leq 1} |\mathrm{CD}_N(t)|\theta(1/l) > \epsilon] \\
&= \mathrm{P}[\sup_{0\leq t\leq 1} |\mathrm{CD}_N(t)| > h^2\epsilon/\theta(1/l)] \\
&\to \mathrm{P}[\sup_{0\leq t\leq 1} |L(t)| > h^2\epsilon/\theta(1/l)] \text{ as } m \wedge n \to \infty \\
&\to 0 \text{ as } l \to \infty.
\end{aligned}$$

The convergence of the probability as $m \wedge n \to \infty$ is a consequence of

$$\sup_{0\leq t\leq 1} |\mathrm{CD}_N(t)| \xrightarrow{\mathrm{d}} \sup_{0\leq t\leq 1} |L(t)|.$$

This convergence in distribution is a result of the weak convergence of the process CD_N to the process L. One need only apply Theorem 3.11 of Ruymgaart (1988) to obtain the convergence in distribution result.

Examine now the last term of (B.9). A consequence of Lemma (B.1) is that KDP_h is uniformly continuous with probability 1. This result implies that $\rho(\mathrm{KDP}_h, A_l(\mathrm{KDP}_h)) \to 0$ as $l \to \infty$ with probability 1. Hence one has the result

$$\mathrm{P}[\rho(\mathrm{KDP}_h, A_l(\mathrm{KDP}_h)) > \epsilon] \to 0 \text{ as } l \to \infty.$$

All that remains to be shown is that

$$\left|\mathrm{E}[g(A_l(\mathrm{KDP}_{N,h}))] - \mathrm{E}[g(A_l(\mathrm{KDP}_h))]\right| \to 0 \text{ as } m \wedge n \to \infty.$$

It can be shown quite easily that $\mathrm{KDP}_{N,h}(w) \xrightarrow{\mathrm{d}} \mathrm{KDP}_h(w)$ for each $w \in [0,1]$. This is shown by appeal to Theorem 3.11 of Ruymgaart (1988). It is similarly shown that for $0 \le w_1 \le w_2 \le \ldots \le w_k \le 1$ and $(b_1, \ldots, b_k) \in \mathbb{R}^k$, one has

$$\sum_{i=1}^{k} b_i \mathrm{KDP}_{N,h}(w_i) \xrightarrow{\mathrm{d}} \sum_{i=1}^{k} b_i \mathrm{KDP}_h(w_i).$$

This result follows from the above convergence in distribution and the fact that integrals are linear operators. By the Cramér-Wold device, one may conclude that

$$\text{(B.10)}\quad (\mathrm{KDP}_{N,h}(w_1), \ldots, \mathrm{KDP}_{N,h}(w_k)) \xrightarrow{\mathrm{d}} (\mathrm{KDP}_h(w_1), \ldots, \mathrm{KDP}_h(w_k))$$

as $m \wedge n \to \infty$. Convergence in distribution implies that

$$\mathrm{E}[h(\mathrm{KDP}_{N,h}(w_1), \ldots, \mathrm{KDP}_{N,h}(w_k))] \to \mathrm{E}[h(\mathrm{KDP}_h(w_1), \ldots, \mathrm{KDP}_h(w_k))]$$

as $m \wedge n \to \infty$ for any bounded and continuous function $h: \mathbb{R}^k \to \mathbb{R}$. Define

$$\phi(x_0, \ldots, x_l) = g(A_l(x)),$$

where $x(s) \in C[0,1]$ and $x_k = x(k/l)$. The function ϕ maps $\mathbb{R}^{l+1}$ into $\mathbb{R}$ since the function A_l depends on only $l+1$ values of the function $x(s)$. Since g is

bounded and continuous in the sup-norm, ϕ is bounded and continuous in the Euclidean norm. Combining (B.10) with these results concerning ϕ yields

$$\begin{aligned}\mathrm{E}[g(A_l(\mathrm{KDP}_{N,h}))] &= \mathrm{E}[\phi(\mathrm{KDP}_{N,h}(0/l), \mathrm{KDP}_{N,h}(1/l), \ldots, \mathrm{KDP}_{N,h}(l/l))] \\ &\to \mathrm{E}[\phi(\mathrm{KDP}_h(0/l), \mathrm{KDP}_h(1/l), \ldots, \mathrm{KDP}_h(l/l))] \\ &= \mathrm{E}[g(A_l(\mathrm{KDP}_h))].\end{aligned}$$

The convergence occurs as $m \wedge n \to \infty$ and holds for all $l \geq 1$. Hence

$$\left|\mathrm{E}[g(A_l(\mathrm{KDP}_{N,h}))] - \mathrm{E}[g(A_l(\mathrm{KDP}_h))]\right| \to 0$$

as $m \wedge n \to \infty$ for all $l \geq 1$.

Combining all these results one has

$$\lim_{l\to\infty} \lim_{m\wedge n\to\infty} \left|\mathrm{E}[g(\mathrm{KDP}_{N,h})] - \mathrm{E}[g(\mathrm{KDP}_h)]\right| = 2\delta_g(\epsilon) \to 0$$

as $\epsilon \to 0$. Thus $\mathrm{KDP}_{N,h} \Rightarrow \mathrm{KDP}_h$ in $(C[0,1], C_\rho, \rho)$ as $m \wedge n \to \infty$.

Proof of Lemma 3.2.3

The covariance kernel, $C_h(w,v)$, of $\sqrt{\lambda_0/(1-\lambda_0)}\mathrm{KDP}_h(w)$ is given by

$$C_h(w,v) = \frac{1}{h^4}\int_0^1 K_s'([w-s]/h)\int_0^1 K_{s'}'([v-t]/h)[\min(s,t) - st]dsdt.$$

See Kannan (1979), page 154, for a proof of this result. Letting

$$f'(s) = \frac{1}{h^2}K_s'([w-s]/h), \tag{B.11}$$

$$g'(t) = \frac{1}{h^2}K_{s'}'([v-t]/h), \tag{B.12}$$

an equivalent expression for $C_h(w,v)$ is

$$\begin{aligned}C_h(w,v) &= \int_0^1 f'(s)\int_0^1 g'(t)[\min(s,t) - st]dsdt \\ &= \int_0^1\int_0^1 f'(s)g'(t)\min(s,t)dsdt - \int_0^1 sf'(s)ds \cdot \int_0^1 tg(t)dt. \qquad \text{(B.13)}\end{aligned}$$

The first term of (B.13) is

$$\begin{aligned}
&\int_0^1 \int_0^1 f'(s)g'(t)\min(s,t)dsdt \\
&= \int_0^1 f'(s) \int_0^s tg'(t)dtds + \int_0^1 sf'(s) \int_s^1 g'(t)dtds \\
&= \int_0^1 f'(s)\Big[tg(t)\Big|_0^s - \int_0^s g(t)dt\Big]ds \\
&\quad + \int_0^1 sf'(s)[g(1) - g(s)]ds \\
&= \int_0^1 f'(s)\Big[sg(s) - \int_0^s g(t)dt\Big]ds \\
&\quad + g(1)\int_0^1 sf'(s)ds - \int_0^1 sg(s)f'(s)ds \\
&= -\int_0^1 \int_0^s f'(s)g(t)dtds + g(1)f(1) - g(1)\int_0^1 f(s)ds \\
&= \int_0^1 g(s)f(s)ds - f(1)\int_0^1 g(t)dt + g(1)f(1) - g(1)\int_0^1 f(s)ds.
\end{aligned}$$

The second term of (B.13) becomes:

$$\begin{aligned}
&\int_0^1 sf'(s)ds \cdot \int_0^1 tg'(t)dt \\
&= \Big[f(1) - \int_0^1 f(s)ds\Big]\Big[g(1) - \int_0^1 g(t)dt\Big].
\end{aligned}$$

There is much cancellation when these are combined to find $C_h(w,v)$. The result is

$$C_h(w,v) = \int_0^1 g(s)f(s)ds - \int_0^1 f(s)ds \cdot \int_0^1 g(t)dt.$$

Defining $f'(s)$ and $g'(t)$ as in (B.11) and (B.12) gives

$$\begin{aligned}
f(s) &= -\frac{1}{h}K_s([w-s]/h), \\
g(t) &= -\frac{1}{h}K_{s'}([v-t]/h).
\end{aligned}$$

The final form for $C_h(w,v)$ is then

$$C_h(w,v) = \frac{1}{h^2}\int_0^1 \int_0^1 K_s([w-s]/h)K_{s'}([v-s]/h)ds - 1.$$

The '-1' arises since the kernel integrates to 1.

It is possible to derive a closed form representation for $C_h(w,v)$. This is very desirable as it is necessary to evaluate $C_h(w,v)$ many times while approximating the eigenvalues and eigenfunctions. The process of finding this representation starts by evaluating the integral:

$$\int_p^q K_s(u)K_{s'}(u+[v-w]/h)du,$$

which results, by a change of variable $u=[w-s]/h$, in the integral

$$\frac{1}{h}\int_x^y K_s([w-s]/h)K_{s'}([v-s]/h)ds, \tag{B.14}$$

where x and y will be determined shortly. Writing out the formula for the boundary kernel, one can actually perform the integration in closed form:

$$\begin{aligned}&\int_p^q K_s(u)K_{s'}([v-w]/h+u)du\\ &=\int_p^q(\theta_s+\phi_s u)(\theta_{s'}+\phi_{s'}(u+[w-v]/h)K(u)K(u+[w-v]/h)du.\end{aligned} \tag{B.15}$$

The kernel $K(u)$ is taken to be the biweight kernel, $K(u)=a(1-u^2)^2$ with $a=15/16$. Substituting this form of K into equation (B.15) and after much simplification, one arrives at the solution

$$\int_p^1 K_s(u)K_{s'}(u+[v-w]/h)du=a^2\sum_{i=0}^{10}\frac{b_i}{i+1}(q^{i+1}-p^{i+1}),$$

where

$$\begin{aligned}b_0&=a_0d,\\ b_1&=a_1d+a_0e,\\ b_2&=a_2d+a_1e+a_0f,\\ b_3&=a_3d+a_2e+a_1f,\\ b_4&=a_4d+a_3e+a_2f,\\ b_5&=a_5d+a_4e+a_3f,\\ b_6&=a_6d+a_5e+a_4f,\end{aligned}$$

$$
\begin{aligned}
b_7 &= a_7 d + a_6 e + a_5 f, \\
b_8 &= a_8 d + a_7 e + a_6 f, \\
b_9 &= a_8 e + a_7 f, \\
b_{10} &= a_8 f, \\
a_0 &= 1 - 2c^2 + c^4, \\
a_1 &= 4c(c^2 - 1), \\
a_2 &= -2(c^4 - 5c^2 + 2), \\
a_3 &= -4c(2c^2 - 3), \\
a_4 &= 6 - 14c^2 + c^4, \\
a_5 &= 4c(c^2 - 3), \\
a_6 &= 2(3c^2 - 2), \\
a_7 &= 4c, \\
a_8 &= 1, \\
c &= (v - w)/h, \\
d &= \theta_s \theta_{s'} + \theta_s \phi_{s'} c, \\
e &= \theta_{s'} \phi_s + \theta_s \phi_{s'} + \phi_s \phi_{s'} c, \\
f &= \phi_s \phi_{s'}.
\end{aligned}
$$

The relation of x and y of equation (B.14) and p and q of equation (B.15) is

$$
\begin{aligned}
p &= \frac{w - y}{h}, \\
q &= \frac{w - x}{h}.
\end{aligned}
$$

The limits x and y need to be determined so that

$$K_s([w - u]/h) K_{s'}([v - u]/h)$$

is non-zero a.e. over this range so that the formulas for the boundary kernel employed to arrive at (B.15) are valid. Recall that the support of the left-hand boundary kernel is $[-1, s]$ and the right-hand boundary kernel is $[-s, 1]$. The integrand will be non-zero outside these intervals so it is very important to limit

the range of integration. The limits are given for the various cases below. Assume without loss of generality that $w \leq v$.

Case 1: $0 \leq w \leq h$ and $0 \leq v \leq h$. Take $x = 0$ and $y = w + h$ which gives $p = -1$ and $q = w/h$.

Case 2: $0 \leq w \leq h$ and $h \leq v \leq 1 - h$.

Case 2a: $w + h \leq v - h$ means the integral is zero.

Case 2b: $w + h > v - h$. Take $x = v - h$ and $y = w + h$ which gives $p = -1$ and $q = 1 - c$.

Case 3: $0 \leq w \leq h$ and $1 - h \leq v \leq 1$.

Case 3a: $w + h \leq v - h$ implies the integral is zero.

Case 3b: $w + h > v - h$. Take $x = v - h$ and $y = w + h$ which gives $p = -1$ and $q = 1 - c$.

Case 4: $h \leq w \leq 1 - h$ and $h \leq v \leq 1 - h$.

Case 4a: $w + h \leq v - h$ implies the integral is zero.

Case 4b: $w + h > v - h$. Take $x = v - h$ and $y = w + h$ which gives $p = -1$ and $q = 1 - c$.

Case 5: $h \leq w \leq 1 - h$ and $1 - h \leq v \leq 1$.

Case 5a: $w + h \leq v - h$ implies the integral is zero.

Case 5b: $w + h > v - h$. Take $x = v - h$ and $y = w + h$ which gives $p = -1$ and $q = 1 - c$.

Case 6: $1 - h \leq w \leq 1$ and $1 - h \leq v \leq 1$. Take $x = v - h$ and $y = 1$ which gives $p = (w - 1)/h$ and $q = 1 - c$.

After implementing these formulas, one finds $C_h(w, v)$ by rewriting it as

$$C_h(w, v) = \frac{1}{h}\Big[\frac{1}{h}\int_x^y K_s([w - u]/h)K_{s'}([v - u]/h)du\Big] - 1.$$

Proof of Lemma 3.3.1

To prove that $C_h(w, v)$ is continuous on the unit square, it must be shown that for all $\epsilon > 0$ there exists a $\delta > 0$ such that

$$\text{(B.16)} \qquad \Big|C_h(w_1, v_1) - C_h(w, v)\Big| < \epsilon,$$

if $\|(w_1, v_1) - (w, v)\| < \delta$ where $\|\cdot\|$ is the Euclidean norm. Equation (B.16) can be rewritten as

$$\begin{aligned}
&\Big|C_h(w_1, v_1) - C_h(w, v)\Big| \\
&= \Big|C_h(w_1, v_1) - C_h(w_1, v) + C_h(w_1, v) - C_h(w, v)\Big| \\
&\leq \sup_{s,t}\Big|\frac{1}{h^2}K_s(t)\Big| \int_0^1 \Big|K([v_1 - u]/h) - K([v - u]/h)\Big| du \\
&\quad + \sup_{s,t}\Big|\frac{1}{h^2}K_s(t)\Big| \int_0^1 \Big|K([w_1 - u]/h) - K([w - u]/h)\Big| du.
\end{aligned} \tag{B.17}$$

With a proof completely analogous to that of Lemma 3.2.1, one can show that

$$\sup_{|w_1 - w| < \delta} \Big|K([w_1 - u]/h) - K([w - u]/h)\Big| du < \theta(\delta),$$

where $\theta(\delta) \to 0$ as $\delta \to 0$. This implies that (B.17) can be made smaller than ϵ if δ is taken sufficiently small and

$$\begin{aligned}
|w_1 - w| &< \delta, \\
|v_1 - v| &< \delta.
\end{aligned}$$

Note that $\|(w_1, v_1) - (w, v)\| < \delta$ implies these last two conditions. It must also be shown that

$$\sup_{s,t}\Big|K_s(t)\Big| < \infty,$$

for any of these bounds to be meaningful. By definition,

$$K_s(t) = (\theta_s + \phi_s t)K(t).$$

For the supremum to be infinite, it is clear that either

$$\sup_s |\theta_s| = \infty$$

or

$$\sup_s |\phi_s| = \infty,$$

since $K(t)$ is bounded and the set over which the supremum is taken with respect to t is also bounded. Without loss of generality, consider the left hand boundary kernel. It is constructed to satisfy

$$\int_{-1}^{s} K_s(t)dt = 1,$$

for all $s \in [0,1]$. Hence the supremum of θ_s and ϕ_s must be bounded.

It has now been shown that $C_h(w,v)$ is continuous on the unit square. In fact, $C_h(w,v)$ is uniformly continuous on the unit square. Continuity implies that $C_h(w,v)$ is bounded and integrable. Hence both integrals given in parts (ii) and (iii) of the lemma are finite.

Proof of Lemma 3.3.2

The results of Lemma 3.3.2 will be proved in reverse order. Since $C_h(w,v)$ is continuous on the unit square, the series

$$C_h(w,v) = \sum_{j=1}^{\infty} \theta_j^h \phi_j^h(w)\phi_j^h(v)$$

converges absolutely and uniformly by Mercer's theorem [see Shorack and Wellner (1986), page 208]. Proposition (ii) of Lemma 3.3.2 is a direct consequence of Lemma 3.3.1 and Proposition 2 of Shorack and Wellner, page 208. Proposition (i) of Lemma 3.3.2 is that the eigenfunctions, $\phi_j^h(w)$, are continuous on $[0,1]$. Start with the series representation given by Mercer's theorem. Since this series converges absolutely, it must be that

(B.18) $$|\phi_j^h(w)\phi_j^h(v)| < \infty,$$

for all $w, v \in [0,1]$. Since $\phi_j^h(w)$ is an eigenfunction, by assumption $\phi_j^h(w) \not\equiv 0$. Let w be such that $\phi_j^h(w) \neq 0$. Combining this result with (B.18) implies the existence of M_j such that

$$|\phi_j^h(v)| \leq M_j < \infty,$$

for all $v \in [0,1]$.

The defining equation for $\phi_j^h(w)$ and θ_j^h leads to

$$\begin{aligned}
&|\phi_j^h(w') - \phi_j^h(w)| \\
&\quad = \frac{1}{\theta_j^h}\Big|\int_0^1 \Big[C_h(v,w') - C_h(v,w)|\phi_j^h(v)dv\Big| \\
&\quad \leq \frac{1}{\theta_j^h}\int_0^1 \Big|C_h(v,w') - C_h(v,w)\Big||\phi_j^h(v)|dv \\
&\quad \leq \frac{\epsilon}{\theta_j^h}\int_0^1 |\phi_j^h(v)|dv, \quad \text{if} \quad |w'-w| < \delta, \\
&\quad \leq \frac{\epsilon M_j}{\theta_j^h}.
\end{aligned}$$

Thus $\phi_j^h(w)$ is continuous on $[0,1]$. The ϵ appears because $C_h(w,v)$ is uniformly continuous on the unit square.

Proof of Lemma 3.3.3

The result of Lemma 3.3.3 was outlined in the text. It remains only to specify some of the details. Choose $(b_1,\ldots,b_M) \in \mathbb{R}^M$ and define

$$\begin{aligned}
X_N &= \sum_{i=1}^M b_i Z_{Ni} \\
&= \int_0^1 \Big[\sum_{i=1}^M b_i \phi_i^h(w)\Big]\mathrm{KDP}_{N,h}(w)dw \\
&\equiv \int_0^1 R(w)\mathrm{KDP}_{N,h}(w)dw,
\end{aligned}$$

where

$$R(w) = \sum_{i=1}^M b_i \phi_i^h(w).$$

Clearly $R(w)$ is continuous since each $\phi_j^h(w)$ is. The functional

$$f(G) = \int_0^1 R(w)G(w)dw$$

is continuous on $(C[0,1],\rho)$ and measurable $(\mathcal{B},\mathcal{C}_\rho)$ (see Ruymgaart (1988), pages 40 ff.). By Theorem 3.11 of Ruymgaart (1988), one can conclude that

$$X_N \xrightarrow{\mathrm{d}} X$$

as $m \wedge n \to \infty$, where X is given by

$$\begin{aligned} X &= \int_0^1 R(w)\mathrm{KDP}_h(w)dw \\ &= \int_0^1 \Big[\sum_{i=1}^M b_i \phi_j^h(w)\Big]\mathrm{KDP}_h(w)dw \\ &= \sum_{i=1}^M b_i \int_0^1 \phi_j^h(w)\mathrm{KDP}_h(w)dw, \end{aligned}$$

since integrals are linear operators. This last equality is clearly $\sum_{i=1}^M b_i Z_i$. It has now been shown that for $(b_1, \ldots, b_M) \in \mathbb{R}^M$ one has

$$\sum_{i=1}^M b_i Z_{Ni} \xrightarrow{\mathrm{d}} \sum_{i=1}^M b_i Z_i.$$

It may be concluded by the Cramér-Wold device that

$$(Z_{N1}, \ldots, Z_{NM}) \xrightarrow{\mathrm{d}} (Z_1, \ldots, Z_M).$$

Proof of Lemma 3.3.4

Let $g(w)$ be any element in S_D^h. The condition for $C_h(w,v)$ to be positive semi-definite is that there exist a $g \in S_D^h$ such that

$$\int_0^1 \int_0^1 g(w)C_h(w,v)g(v)dwdv = 0.$$

The integral can be rewritten in the following manner:

$$\begin{aligned} &\int_0^1 \int_0^1 g(w)C_h(w,v)g(v)dwdv \\ &= \int_0^1 \int_0^1 g(w)\Big[\frac{1}{h^2}\int_0^1 K_s([w-u]/h) \\ &\qquad \cdot K_{s'}([v-u]/h)du - 1\Big]g(v)dwdv \\ &= \int_0^1 \Big[\frac{1}{h}\int_0^1 g(w)K_s([w-u]/h)dw\Big] \\ &\qquad \cdot \Big[\frac{1}{h}\int_0^1 g(v)K_{s'}([v-u]/h)dv\Big]du - \Big[\int_0^1 g(t)dt\Big]^2, \end{aligned} \tag{B.19}$$

where $s = s(w,h)$ and $s' = s(v,h)$. Let

$$r(u) = \frac{1}{h}\int_0^1 g(w)K_s([w-u]/h)dw,$$

and note that

$$\int_0^1 r(u)du = \int_0^1 g(t)dt.$$

Substituting $r(u)$ into equation (B.19), one obtains

$$\begin{aligned}&\int_0^1\int_0^1 g(w)C_h(w,v)g(v)dwdv\\ &= \int_0^1 r(u)^2du - \mu^2\\ &= \int_0^1 [r(u)-\mu]^2du,\end{aligned}$$

where

$$\mu = \int_0^1 r(u)du.$$

If $r(u) = c$ then

$$\int_0^1 [r(u)-\mu]^2du = \int_0^1 [c-c]^2du = 0.$$

Conversely, if

$$\int_0^1 [r(u)-\mu]^2du = 0 \tag{B.20}$$

then

$$r(u) = c. \tag{B.21}$$

This follows since the integrand in (B.20) is non-negative and $r(u)$ is continuous. Putting these results together one finds the following two results.

1. If there exists a $g \in S_D^h$, $g \not\equiv 0$, such that $r(u) = c$ then (B.20) holds and thus

$$\int_0^1\int_0^1 g(w)C_h(w,v)g(v)dwdv = 0. \tag{B.22}$$

2. If there exists $g \in S_D^h$, $g \not\equiv 0$, such that (B.22) holds then (B.20) also holds and hence $r(u) = c$.

Thus there exists a $g \in S_D^h$, $g \not\equiv 0$, such that (B.22) holds if and only if there exists a $g \in S_D^h$, $g \not\equiv 0$ such that (B.21) holds.

It may be concluded that the condition that there exist a $g \in S_D^h$ such that

$$\frac{1}{h}\int_0^1 K_s([w-u]/h)g(w)dw = c, \quad \forall u \in [0,1]$$

is equivalent to the condition for positive semi-definiteness.

Proof of Lemma 3.3.5

The subset chi-square test depends on the data only through the components. The components are invariant when centered by the small sample mean. Hence, the chi-square test is invariant.

Proof of Lemma 3.3.6

The invariance of the orthogonal series estimator also results from the invariance of the components. Lemma 3.3.5 states that the set of components selected by the subset chi-square test will be the same irrespective of which sample is called the first. This result is due to the invariance of the subset chi-square test applied to the components. Let $\hat{d}_{h,M}^F$ be the orthogonal series estimate and U_{Ni}^* be the unnormalized components when the population with distribution function F is termed the first sample. Let $\hat{d}_{h,M}^G$ be the orthogonal series estimate and V_{Ni}^* be the unnormalized components when the population with distribution function G is termed the first sample. Let $\lambda_{(N)} = m/N$ be the proportion of the total sample represented by the population with distribution function F.

The claim is that

$$\lambda_{(N)}\hat{d}_{h,M}^F(w) + (1-\lambda_{(N)})\hat{d}_{h,M}^G(w) = 1$$

for all $w \in [0,1]$. Let the normalized components be U_{Ni} and V_{Ni}. They are

given by

$$U_{Ni} = \sqrt{\frac{N\lambda_{(N)}}{\theta_i^h(1-\lambda_{(N)})}}U_{Ni}^*$$
$$= \sqrt{\frac{Nm}{n\theta_i^h}}U_{Ni}^*,$$

and

$$V_{Ni} = \sqrt{\frac{N(1-\lambda_{(N)})}{\theta_i^h\lambda_{(N)}}}V_{Ni}^*$$
$$= \sqrt{\frac{Nn}{m\theta_i^h}}V_{Ni}^*.$$

The invariance condition for the components is $U_{Ni} = -V_{Ni}$. In terms of U_{Ni}^* and V_{Ni}^* this is

(B.23) $$mU_{Ni}^* = -nV_{Ni}^*.$$

Let S represent the set of components in the models and write out the invariance condition:

$$\lambda_{(N)}\hat{d}_{h,M}^F(w) + (1-\lambda_{(N)})\hat{d}_{h,M}^G(w)$$
$$= \frac{m}{N}\Big[1 + \sum_{i\in S}U_{Ni}^*\phi_i^h(w)\Big] + \frac{n}{N}\Big[1 + \sum_{i\in S}V_{Ni}^*\phi_i^h(w)\Big]$$
$$= 1 + \frac{1}{N}\sum_{i\in S}(mU_{Ni}^* + nV_{Ni}^*)\phi_i^h(w)$$
$$= 1,$$

in light of equation (B.23). Thus, the orthogonal series estimator is invariant.

Proof of Theorem 4.2.1

The proof of this theorem will make heavy use of the theorems and techniques of Pyke and Shorack (1968). In fact, the weak convergence will be shown for their process. Since the process CDo_N is asymptotically equivalent to their process (see Section 2), it will inherit the result.

It must be shown that

$$\begin{aligned}&\|Lo_N(t) - L_o(t) - (1-\lambda_0)^{-1/2}\Delta(t)\| \\ &\leq \|L_N(t) - L_o(t)\| + \|\sqrt{N}[D^F_{(N)}(t) - t] - (1-\lambda_0)^{-1/2}\Delta(t)\| \xrightarrow{\mathrm{P}} 0,\end{aligned}$$

as $m \wedge n \to \infty$, where

$$\begin{aligned}Lo_N(t) &= \sqrt{N}[F_m Q^H_N(t) - t], \\ L_N(t) &= \sqrt{N}[F_m Q^H_N(t) - D^F_{(N)}(t)], \\ L_o(t) &= \sqrt{\frac{1-\lambda_0}{\lambda_0}} B(t) \\ H_{(N)}(x) &= \lambda_{(N)} F(x) + (1-\lambda_{(N)}) G_{(n)}(x), \\ Q^H_{(N)}(t) &= H^{-1}_{(N)}(t), \\ D^F_{(N)}(t) &= F Q^H_{(N)}(t),\end{aligned}$$

$B(t)$ is a Brownian bridge and $\|\cdot\|$ denotes the sup-norm. By assumption, one has the result

$$\|\sqrt{N}[D^F_{(N)}(t) - t] - (1-\lambda_0)^{-1/2}\Delta(t)\| \to 0,$$

as $m \wedge n \to \infty$.

Start by giving an alternate representation of the Pyke-Shorack process, $L_N(t)$, in the Skorohod space

$$\begin{aligned}L_N(t) = (1-\lambda_{(N)})\Big\{ &\frac{1}{\sqrt{\lambda_{(N)}}} B_N(t) U_m[F Q^H_N(t)] \\ &- \frac{1}{\sqrt{1-\lambda_{(N)}}} A_N(t) V_n[G_{(n)} Q^H_N(t)] \Big\} + \delta_N(t),\end{aligned}$$

where

$$\begin{aligned}\delta_N(t) &= A_N(t)\sqrt{N}[H_N Q^H_N(t) - t], \\ A_N(t) &= [D^F_{(N)}(u_t) - D^F_{(N)}(t)]/(u_t - t), \\ u_t &= H_{(N)} Q^H_N(t),\end{aligned}$$

$$\lambda_{(N)}A_N(t) + (1-\lambda_{(N)})B_N(t) = 1,$$

$$|A_N(t)| \leq 1/\lambda_{(N)},$$

$$|B_N(t)| \leq 1/(1-\lambda_{(N)}),$$

$$|\delta_N| \leq 1/(\lambda_{(N)}\sqrt{N}),$$

$$D^G_{(N)}(t) = G_{(n)}Q^H_{(N)}(t),$$

$$U_m(t) = \sqrt{m}[F_mQ^F(t) - t],$$

$$V_n(t) = \sqrt{n}[G_nQ^G_{(n)}(t) - t].$$

It will be shown that $\sup_{0\leq t\leq 1}|L_N(t) - L_o(t)| \xrightarrow{\text{p}} 0$ as $m \wedge n \to \infty$. The process L_o is the limiting process of L_N under H_o.

The proof divides the interval $[0,1]$ into three subintervals: $[0,1/N]$, $[1/N, 1 - 1/N]$, and $[1-1/N, 1]$. For the first and last intervals, the goal is to show that

$$\sup|L_N(t)| \xrightarrow{\text{p}} 0,$$
$$\sup|L_o(t)| \xrightarrow{\text{p}} 0,$$

as $m \wedge n \to \infty$. Start by examining the $L_N(t)$ on the first interval:

$$\sup_{0\leq t\leq 1/N}|L_N(t)| \leq \sup_{0\leq t\leq 1/N}\frac{\sqrt{N}}{m} + \sqrt{N}D^F_{(N)}(t),$$

since

$$L_N(t) \equiv \sqrt{N}[F_mQ^H_N(t) - D^F_{(N)}(t)],$$

and

$$F_mQ^H_N(t) \leq \frac{1}{m} \text{ on } 0 \leq t \leq \frac{1}{N}.$$

Bound $D^F_{(N)}(t)$ on $[0,1/N]$ in the following manner:

$$\lambda_{(N)}D^F_{(N)}(t) + (1-\lambda_{(N)})D^G_{(N)}(t) = t$$

for $0 \leq t \leq 1$, so

$$\lambda_{(N)}D^F_{(N)}(t) = t - (1-\lambda_{(N)})D^G_{(N)} \leq t$$

and

$$D^F_{(N)}(t) \leq \frac{t}{\lambda_{(N)}} \leq \frac{1}{N\lambda_{(N)}}$$

on $[0, 1/N]$. Hence

$$0 \leq \sup_{0\leq t\leq 1/N} |L_N(t)| \leq \frac{\sqrt{N}}{m} + \frac{1}{\sqrt{N}\lambda_{(N)}} \to 0.$$

Similarly, it can be shown that

$$\sup_{1-1/N\leq t\leq 1} |L_N(t)| \to 0.$$

Equation (2.5) of Aly, Csörgő, and Horváth (1987) gives

$$\lim_{\epsilon\to 0^+} \mathrm{P}[\sup_{0\leq t\leq \epsilon} |B(t)| \geq \delta] = 0,$$
$$\lim_{\epsilon\to 0^+} \mathrm{P}[\sup_{1-\epsilon\leq t\leq 1} |B(t)| \geq \delta] = 0,$$

for all $\delta > 0$. Let $\epsilon = 1/N$ and choose $\delta > 0$, then

$$\begin{aligned}
&\mathrm{P}[\sup_{0\leq t\leq 1/N} |L_N(t) - L_o(t)| \geq \delta] \\
&\leq \mathrm{P}[\sup_{0\leq t\leq 1/N} |L_N(t)| + \sup_{0\leq t\leq 1/N} |L_o(t)| \geq \delta] \\
&\leq \mathrm{P}[\sup_{0\leq t\leq 1/N} |L_N(t)| \geq \delta/2] + \mathrm{P}[\sup_{0\leq t\leq 1/N} |L_o(t)| \geq \delta/2] \\
&\to 0 \text{ as } m \wedge n \to \infty.
\end{aligned}$$

The procedure for the interval $[1-1/N, 1]$ is perfectly analogous. This leaves the interval $[1/N, 1-1/N]$. The limiting process, $L_o(t)$, can be represented by

$$L_o(t) = (1-\lambda_0)\Big[\frac{1}{\sqrt{\lambda_0}}U_o(t) - \frac{1}{\sqrt{1-\lambda_0}}V_o(t)\Big],$$

where $U_o(t)$ and $V_o(t)$ are independent Brownian bridges and are the limiting processes of $U_m(t)$ and $V_n(t)$, respectively. Consider the following inequality:

$$\sup_{1/N\leq t\leq 1-1/N} |L_N(t) - L_o(t)|$$

$$
\begin{aligned}
&= \sup\Big|(1-\lambda_{(N)})\Big\{\frac{1}{\sqrt{\lambda_{(N)}}}B_N(t)U_m[FQ_N^H(t)] \\
&\quad - \frac{1}{\sqrt{1-\lambda_{(N)}}}A_N(t)V_n[G_{(n)}Q_N^H(t)] + \delta_N(t)\Big\} \\
&\quad - (1-\lambda_0)\Big\{\frac{1}{\sqrt{\lambda_0}}U_o(t) - \frac{1}{\sqrt{1-\lambda_0}}V_o(t)\Big\}\Big|, \\
&\le R_{1N} + R_{2N} + R_{3N} + R_{4N} + R_{5N},
\end{aligned}
$$

where

$$
\begin{aligned}
R_{1N} &= \sup\Big|\Big[\frac{1-\lambda_{(N)}}{\sqrt{\lambda_{(N)}}} - \frac{1-\lambda_0}{\sqrt{\lambda_0}}\Big]B_N(t)U_m[FQ_N^H(t)]\Big|, \\
R_{2N} &= \sup\Big|\frac{1-\lambda_0}{\sqrt{\lambda_0}}\Big[B_N(t)U_m[FQ_N^H(t)] - U_o(t)\Big]\Big|, \\
R_{3N} &= \sup|\delta_N(t)|, \\
R_{4N} &= \sup\Big|\Big[\sqrt{1-\lambda_0} - \sqrt{1-\lambda_{(N)}}\Big]A_N(t)V_n[G_{(n)}Q_N^H(t)]\Big|, \\
R_{5N} &= \sup\Big|\sqrt{1-\lambda_0}\Big[A_N(t)V_n[G_{(n)}Q_N^H(t)] - V_o(t)\Big]\Big|.
\end{aligned}
$$

The suprema are taken over the range $1/N \le t \le 1 - 1/N$. Henceforth, if there is no range indicated on a supremum, it is assumed to be over the interval $[1/N, 1-1/N]$. Each of the terms R_{1N} through R_{5N} must be shown to tend to zero in probability as $m \wedge n \to \infty$. Terms R_{1N} and R_{4N} tend to zero in probability because $\lambda_{(N)} \to \lambda_0$, U_m and V_n are each bounded in probability (their suprema actually converge in distribution to proper random variables), and $A_N(t)$ and $B_N(t)$ are bounded (see above). Now examine term R_{2N} in detail:

$$
\begin{aligned}
&\sup\Big|B_N(t)U_m[FQ_N^H(t)] - U_o(t)\Big| \\
&\le S_{1N} + S_{2N} + S_{3N},
\end{aligned}
$$

where

$$
\begin{aligned}
S_{1N} &= \sup|B_N(t)| \cdot \Big|U_m[FQ_N^H(t)] - U_o[FQ_{(N)}^H(t)]\Big|, \\
S_{2N} &= \sup|B_N(t)| \cdot \Big|U_o[FQ_{(N)}^H(t)] - U_o(t)\Big|, \\
S_{3N} &= \sup|U_o(t)| \cdot |B_N(t) - 1|.
\end{aligned}
$$

The term $B_N(t)$ is bounded as given above. Theorem 2.2 of Pyke and Shorack states that

$$\sup\left|U_m[FQ_N^H(t)] - U_o[FQ_{(N)}^H(t)]\right| \xrightarrow{\mathrm{P}} 0$$

as $m \wedge n \to \infty$ uniformly in all continuous F, G and $0 < \lambda_{(N)} < 1$. So it matters not that $G(x) = G_{(r.)}(x)$. Thus term S_{1N} converges in probability to zero.

Since $\|FQ_{(N)}^H(t) - t\| \to 0$ (by Poiyá's theorem) and U_o is uniformly continuous almost surely, the term S_{2N} converges in probability to zero.

The term S_{3N} is all that remains. Rewrite $B_N(t)$ as

$$B_N(t) = \frac{D_{(N)}^G[H_{(N)}Q_N^H(t)] - D_{(N)}^G(t)}{H_{(N)}Q_N^H(t) - t}.$$

The mean value theorem implies the existence, for each $t \in [1/N, 1 - 1/N]$ of $s = s_N(t)$ between t and $H_{(N)}Q_N^H(t)$ such that

$$\begin{aligned} B_N(t) &= d_{(N)}^G(s) \\ &= \frac{g_{(n)}[Q_{(N)}^H(s)]}{\lambda_{(N)} f[Q_{(N)}^H(s)] + (1 - \lambda_{(N)}) g_{(n)}[Q_{(N)}^H(s)]} \\ &= \frac{1}{1 - \lambda_{(N)} + \lambda_{(N)} f(u)/g_{(n)}(u)}, \end{aligned} \tag{B.24}$$

where $u = Q_{(N)}^H(s) = Q_{(N)}^H[s_N(t)]$ and is between $Q_{(N)}^H(t)$ and $Q_N^H(t)$. Define the event E_N to be

$$E_N = \left\{a_N \le Q_N^H(t) \le b_N \text{ for } 1/N \le t \le 1 - 1/N\right\}$$

and E_N^c to be its complement. By assumption

$$\mathrm{P}[E_N] \to 1 \text{ as } m \wedge n \to \infty. \tag{B.25}$$

For $\delta > 0$:

$$\begin{aligned} &\mathrm{P}[\sup|B_N(t) - 1| > \delta] \\ &= \mathrm{P}[\sup|B_N(t) - 1| > \delta | E_N] \cdot \mathrm{P}[E_N] \\ &\quad + \mathrm{P}[\sup|B_N(t) - 1| > \delta | E_N^c] \cdot \mathrm{P}[E_N^c]. \end{aligned} \tag{B.26}$$

Consider $\sup|B_N(t) - 1|$ given that E_N holds. Using (B.24) one can rewrite $|B_N(t) - 1|$ as

$$|B_N(t) - 1| = \left| \frac{\lambda_{(N)}[1 - f(u)/g_{(n)}(u)]}{1 - \lambda_{(N)} + \lambda_{(N)} f(u)/g_{(n)}(u)} \right|.$$

Recall that $u = u_N(t)$ is between $Q^H_{(N)}(t)$ and $Q^H_N(t)$ so that

$$e_N \leq u_N(t) \leq f_N \text{ for } 1/N \leq t \leq 1 - 1/N,$$

where e_N and f_N are as defined in the statement of Theorem 4.2.1. Hence, one has

$$\sup|B_N(t) - 1| \leq \frac{\lambda_{(N)} \sup_{e_N \leq x \leq f_N} |1 - f(x)/g_{(n)}(x)|}{1 - \lambda_{(N)} + \lambda_{(N)} \inf_{e_N \leq x \leq f_N} f(x)/g_{(n)}(x)} \longrightarrow 0$$

as $m \wedge n \to \infty$ since by assumption,

$$\sup_{e_N \leq x \leq f_N} f(x)/g_{(n)}(x) \to 1,$$

and

$$\inf_{e_N \leq x \leq f_N} f(x)/g_{(n)}(x) \to 1,$$

as $m \wedge n \to \infty$. Hence, it has been shown that

$$\mathrm{P}[\sup_{1/N \leq t \leq 1-1/N} |B_N(t) - 1| > \delta | E_N] \to 0.$$

Returning to equation (B.26), it is concluded that

$$\mathrm{P}[\sup_{1/N \leq t \leq 1-1/N} |B_N(t) - 1| > \delta] \to 0$$

in light of (B.25) and the above result. Since $\sup_{0 \leq t \leq 1} |U_o(t)|$ is a proper random variable (in fact, it is the limiting distribution of the KS statistic), term S_{3N} tends to zero in probability. Term R_{5N} behaves in an analogous fashion. Hence it has been proved that

$$\|L_{ON}(t) - L_o(t) - (1 - \lambda_0)^{-1/2} \Delta(t)\| \xrightarrow{\mathrm{p}} 0,$$

in the Skorohod construction and so weak convergence results for the original process.

Proof of Lemma 4.2.1

Lemma 4.2.1 gives the uniform convergence of $\sqrt{n}[D_{(N)}(u) - u]$ to $\Delta(u)$ for location and scale alternatives. The complete proof will be given for location alternatives. Changes necessary for scale alternatives will be indicated at the end. Recall, by assumption, that $\lambda_{(N)} = \lambda_0$, hence $D_{(N)}(u) = D_{(n)}(u)$. The n enters because of the local alternative, m does not enter. Start the proof by showing some useful facts:

$$\begin{aligned}\sqrt{n}[D_{(n)}(u) - u] &= \sqrt{n}[FQ^H_{(n)}(u) - u] \\ &= \sqrt{n}[FQ^H_{(n)}(u) - H_{(n)}(u)Q^H_{(n)}(u)] \\ &= \sqrt{n}(1-\lambda_0)\Big[FQ^H_{(n)}(u) - F[Q^H_{(n)}(u) - \gamma/\sqrt{n}]\Big] \\ &= (1-\lambda_0)\gamma f(c), \qquad \text{(B.27)}\end{aligned}$$

where $c = c_n(u)$ is between $Q^H_{(n)}(u)$ and $Q^H_{(n)}(u) - \gamma/\sqrt{n}$. Next it will be shown that $Q^H_{(n)}(u) \downarrow Q^F(u)$ on $[\delta, 1-\delta]$ for each $0 < \delta < 1/2$. Start with the identity

$$\begin{aligned}u &\equiv H_{(n)}Q^H_{(n)}(u) \\ &\equiv \lambda_0 FQ^H_{(n)}(u) + (1-\lambda_0)F[Q^H_{(n)}(u) - \gamma/\sqrt{n}].\end{aligned}$$

Differentiate this last identity with respect to n:

$$\begin{aligned}0 &= \lambda_0 f Q^H_{(n)}(u)\frac{d}{dn}Q^H_{(n)}(u) \\ &\quad + (1-\lambda_0)f[Q^H_{(n)}(u) - \gamma/\sqrt{n}]\Big(\frac{d}{dn}Q^H_{(n)}(u) + \gamma/(2n^{1.5})\Big).\end{aligned}$$

The resulting formula for $(d/dn)Q^H_{(n)}(u)$ is

$$\frac{d}{dn}Q^H_{(n)}(u) = -\frac{\Big[(1-\lambda_0)\gamma/(2n^{1.5})\Big]f[Q^H_{(n)}(u) - \gamma/\sqrt{n}]}{\lambda_0 f Q^H_{(n)}(u) + (1-\lambda_0)f[Q^H_{(n)}(u) - \gamma/\sqrt{n}]} < 0.$$

Thus $Q^H_{(n)}(u) \downarrow Q^F(u)$ for each $u \in [\delta, 1-\delta]$ for $0 < \delta < 1/2$. Since $Q^F(u)$ is continuous on $[\delta, 1-\delta]$, conclude by Dini's theorem that $Q^H_{(n)}$ converges uniformly to $Q^F(u)$ on $[\delta, 1-\delta]$.

The preliminaries are now taken care of and the proof of the lemma may begin. Choose δ such that $0 < \delta < 1/2$ and break the interval $[0,1]$ into $[0,\delta]$, $[\delta, 1-\delta]$, and $[1-\delta, 1]$. Now write

$$\begin{aligned}
&\sup_{0\le u\le 1}\left|\sqrt{n}[D_{(n)}(u) - u] - \Delta(u)\right| \\
&\le A_{1n}(\delta) + A_{2n}(\delta) + A_{3n}(\delta) + A_{4n}(\delta) + A_{5n}(\delta),
\end{aligned}$$

where

$$\begin{aligned}
A_{1n}(\delta) &= \sup_{0\le u\le \delta}\left|\sqrt{n}[D_{(n)}(u) - u]\right|, \\
A_{2n}(\delta) &= \sup_{0\le u\le \delta}|\Delta(u)|, \\
A_{3n}(\delta) &= \sup_{\delta\le u\le 1-\delta}\left|\sqrt{n}[D_{(n)}(u) - u] - \Delta\right|, \\
A_{4n}(\delta) &= \sup_{1-\delta\le u\le 1}\left|\sqrt{n}[D_{(n)}(u) - u]\right|, \\
A_{5n}(\delta) &= \sup_{1-\delta\le u\le 1}|\Delta(u)|.
\end{aligned}$$

Clearly,

$$\lim_{n\to\infty}\sup_{0\le u\le 1}\left|\sqrt{n}[D_{(n)}(u) - u] - \Delta(u)\right| \le \lim_{n\to\infty}\sum_{i=1}^{5} A_{in}(\delta),$$

for all $\delta > 0$ and so

$$\lim_{n\to\infty}\sup_{0\le u\le 1}\left|\sqrt{n}[D_{(n)}(u) - u] - \Delta(u)\right| \le \lim_{\delta\to 0^+}\lim_{n\to\infty}\sum_{i=1}^{5} A_{in}(\delta).$$

The strategy will be to evaluate the limits on the right of this last inequality.

Start with $A_{1n}(\delta)$:

$$\sup_{0\le u\le \delta}\left|\sqrt{n}[D_{(n)}(u) - u]\right| = \sup_{0\le u\le \delta}\left|(1-\lambda_0)\gamma f(c)\right|,$$

from (B.27), and so

$$\sup_{0\le u\le \delta}\left|\sqrt{n}[D_{(n)}(u) - u]\right| \le \sup_{0\le u\le \delta}(1-\lambda_0)\gamma f Q^{H}_{(n)}(u),$$

since $c = c_N(u) \le Q_N^H(u)$ and $fQ_{(n)}^H(u) \to 0$ as $u \to 0^+$. Then

$$\sup_{0\le u\le\delta}\left|\sqrt{n}[D_{(n)}(u)-u]\right| \le \sup_{0\le u\le\delta}(1-\lambda_0)\gamma fQ_{(1)}^H(u),$$

since $Q_1^H(u) \ge Q_{(n)}^H(u)$ and $fQ_{(1)}^H(u) \to 0$ as $u \to 0^+$. Thus,

$$\begin{aligned}\lim_{\delta\to 0^+}\lim_{n\to\infty}\left|\sqrt{n}[D_{(n)}(u)-u]\right| &\le \lim_{\delta\to 0^+}\lim_{n\to\infty}(1-\lambda_0)\gamma\sup_{0\le u\le\delta} fQ_{(1)}^H(u)\\ &= \lim_{\delta\to 0^+}(1-\lambda_0)\gamma\sup_{0\le u\le\delta} fQ_{(1)}^H(u) = 0,\end{aligned}$$

since $\lim_{u\to 0^+} fQ^F(u) = 0$ and f is continuous.

For A_{2n} one has

$$\lim_{\delta\to 0^+}\lim_{n\to\infty}\sup_{0\le u\le\delta}|\Delta(u)| = 0,$$

since Δ is continuous and $\Delta(0) = 0$.

For A_{3n}, one has by equation (B.27) that

$$\sup_{\delta\le u\le 1-\delta}\left|\sqrt{n}[D_{(n)}(u)-u]-\Delta(u)\right| = \sup_{\delta\le u\le 1-\delta}\left|(1-\lambda_0)\gamma[f(c)-fQ^F(u)]\right|.$$

Expand $f(c)$ about $Q^F(u)$ to arrive at

$$\sup_{\delta\le u\le 1-\delta}\left|\sqrt{n}[D_{(n)}(u)-u]-\Delta(u)\right| = \sup_{\delta\le u\le 1-\delta}\left|(1-\lambda_0)\gamma f'(d)[c-Q^F(u)]\right|,$$

where $d = d_n(u)$ is between $c = c_n(u)$ and $Q^F(u)$. One can write

$$\begin{aligned}|c-Q^F(u)| &\le \max\left(\left|Q^F(u)-Q_{(n)}^H(u)\right|, \left|Q^F(u)-Q_{(n)}^H(u)-\gamma/\sqrt{n}\right|\right)\\ &\le \left|Q^F(u)-Q_{(n)}^H(u)\right|+\gamma/\sqrt{n},\end{aligned}$$

since c is between $Q_{(n)}^H$ and $Q_{(n)}^H - \gamma/\sqrt{n}$. Substituting this result in the above formula yields

$$\begin{aligned}&\sup_{\delta\le u\le 1-\delta}\left|\sqrt{n}[D_{(n)}(u)-u]-\Delta(u)\right|\\ &\le \sup_{\delta\le u\le 1-\delta}(1-\lambda_0)\gamma|f'(d)|\cdot\left[\left|Q^F(u)-Q_{(n)}^H(u)\right|+\gamma/\sqrt{n}\right] \longrightarrow 0,\end{aligned}$$

as $n \to \infty$ for all $0 < \delta < 1/2$ since f' is bounded and $Q^H_{(n)}$ converges uniformly to Q^F on all intervals of this form. Hence

$$\lim_{\delta \to 0^+} \lim_{n \to \infty} \sup_{\delta \le u \le 1-\delta} \left| \sqrt{n}[D_{(n)}(u) - u] - \Delta(u) \right| = \lim_{\delta \to 0^+} 0 = 0.$$

The terms A_{4n} and A_{5n} behave the same as A_{1n} and A_{2n}, respectively.

The procedure for scale alternatives is very similar. In this case, one finds that

$$\frac{d}{dn} Q^H_{(n)}(u) = -\frac{(1-\lambda_0)\gamma f[Q^H_{(n)}/(1+\gamma/\sqrt{n})]Q^H_{(n)}(u)/[2n^{1.5}(1+\gamma/\sqrt{n})^2]}{\lambda_0 f Q^H_{(n)}(u) + (1-\lambda_0) f[Q^H_{(n)}(u)/(1+\gamma/\sqrt{n})]/(1+\gamma/\sqrt{n})},$$

which has at most one sign change so that the convergence is still uniform on $[\delta, 1-\delta]$. In this case, the mean value theorem gives

$$\sqrt{n}[D_{(n)}(u) - u] = (1-\lambda_0)\gamma f(c) Q^H_{(n)}(u)/(1+\gamma/\sqrt{n}),$$

where $c = c_n(u)$ is between $Q^H_{(n)}(u)$ and $Q^H_{(n)}(u)/(1+\gamma/\sqrt{n})$. The term $\sqrt{n}[D_{(n)}(u) - u] - \Delta(u)$ can be written as

$$\begin{aligned}(1-\lambda_0)\gamma \Big[& f(c)[Q^H_{(n)}(u) - Q^F(u)]/(1+\gamma/\sqrt{n}) \\ & + Q^F(u)[f(c) - f Q^F(u)]/(1+\gamma/\sqrt{n}) \\ & + f Q^F(u) Q^F(u)[1/(1+\gamma/\sqrt{n}) - 1]\Big].\end{aligned}$$

The intervals $[0, \delta]$ and $[1-\delta, 1]$ are handled as before. The interval $[\delta, 1-\delta]$ uses the result just above and the uniform convergence of $Q^H_{(n)}$ to Q^F on $[\delta, 1-\delta]$. Hence, the convergence of $\sqrt{n}[D_{(n)}(u) - u]$ to $\Delta(u)$ is uniform for both location and scale alternatives satisfying the conditions of Lemma 4.2.1.

Proof of Theorem 4.2.2

Define

$$g(t,x) = \cos(tx)\mathrm{Re}[\phi(t)] + \sin(tx)\mathrm{Im}[\phi(t)],$$

$$\hat{f}_{N,M}(x) = \frac{M}{\pi N} \sum_{j=1}^{N} g(-M + 2M(j-1)/N, x),$$

$$\hat{F}_{N,M}(x_k) = \frac{\pi}{M}\Big[\sum_{j=1}^{k} \hat{f}_{N,M}(x_j) - \frac{1}{2}\hat{f}_{N,M}(x_1) - \frac{1}{2}\hat{f}_{N,M}(x_k)\Big],$$

$$\text{for } x_k = \frac{\pi k}{M}, \quad k = 0, \ldots, [N/2],$$

$$F_{N,M}(x) = \int_0^x \hat{f}_{N,M}(t)dt.$$

The function $\hat{F}_{N,M}(b)$ is defined by linear interpolation if b cannot be written in the form $\pi k/M$, for some integer k between 0 and $[N/2]$.

The proof that $\hat{F}_{N,M}(b) \to F(b)$ will be divided into two parts. First rewrite $\hat{F}_{N,M}(b) - F(b)$ as

$$\hat{F}_{N,M}(b) - F(b) = \Big[\hat{F}_{N,M}(b) - F_{N,M}(b)\Big] + \Big[F_{N,M}(b) - F(b)\Big].$$

Each of the two terms will be shown to tend to zero.

Define x_b to be the nearest x less than or equal to b such that

$$x_b = \frac{\pi k}{M},$$

for some integer k, $0 \le k \le [N/2]$. Let x_a be the next greatest x of this form:

$$x_a = \pi\frac{k+1}{M}.$$

Clearly $x_a \downarrow b$ and $x_b \uparrow b$ as $M, N \to \infty$. Since $\hat{F}_{N,M}(x_b)$ is approximating $F_{N,M}(x_b)$ by the trapezoidal rule, one has the bound [see Press, Flannery, Teukolsky, and Vetterling (1986), page 105]

$$\begin{aligned}
&\Big|\hat{F}_{N,M}(x_b) - F_{N,M}(b)\Big| \\
&\le \Big|\hat{F}_{N,M}(x_b) - F_{N,M}(x_b)\Big| + \Big|F_{N,M}(x_b) - F_{N,M}(b)\Big| \\
&\le \sup_{0\le x\le b}\Big|\frac{d^2}{dx^2}\hat{f}_{M,N}(x)\Big| O(b/M^2) + |\hat{f}_{N,M}(b)(x_b - b)| + o(x_b - b), \qquad \text{(B.28)}
\end{aligned}$$

since $k \sim bM/\pi$ for large M and the trapezoidal bound is

$$O(a^3/n^2) \sup_{0\le x\le a} |f''(x)|$$

when integrating $f(x)$ over the range $[0,a]$ with n grid points. Young's form of Taylor's theorem is used to derive the last two terms of (B.28). The term $\hat{f}_{N,M}(b)$ can be seen to converge to $f(b)$ by the results below and so

$$|\hat{f}_{N,M}(b)(x_b - b)| \to 0$$

as $N, M \to \infty$ and $M/N \to 0$. The same result holds for $\hat{F}_{N,M}(x_a)$. The second derivative with respect to x of $\hat{f}_{N,M}(x)$ can be bounded:

$$\left|\frac{d^2}{dx^2}\hat{f}_{N,M}(x)\right| = \left|\frac{M}{\pi N}\sum_{j=1}^{N}\frac{\partial^2}{\partial x^2}g(t,x)\right|$$

$$\leq \frac{1}{\pi}\Big[\frac{M}{N}\sum_{j=1}^{N}\left|t_j^2\mathrm{Re}[\phi(t_j)]\right| + \frac{M}{N}\sum_{j=1}^{N}\left|t_j^2\mathrm{Im}[\phi(t_j)]\right|\Big],$$

where $t_j = -M + 2(j-1)M/N$. The expression in the brackets is converging to

$$\int_{-\infty}^{\infty}\left|t^2\mathrm{Re}[\phi(t)]\right|dt + \int_{-\infty}^{\infty}\left|t^2\mathrm{Im}[\phi(t)]\right|dt,$$

as $M, N \to \infty$, and $M/N \to 0$ and so is bounded since these integrals are by assumption. One may conclude that

$$\hat{F}_{N,M}(x_a) - F_{N,M}(b) \to 0,$$
$$\hat{F}_{N,M}(x_b) - F_{N,M}(b) \to 0,$$

as $M, N \to \infty$ and $M/N \to 0$ because the differences are $O(b/M^2)$. Since $\hat{F}_{N,M}(b)$ is between $\hat{F}_{N,M}(x_b)$ and $\hat{F}_{N,M}(x_a)$, it may be concluded that

$$\hat{F}_{N,M}(b) - F_{N,M}(b) \to 0,$$

as $N, M \to \infty$ and $M/N \to 0$.

Next it will be shown that $F_{N,M}(b) - F(b)$ is tending to zero. Write

$$F_{N,M}(b) - F(b) = \int_0^b [\hat{f}_{N,M}(x) - f(x)]dx$$
$$= \int_0^b [\hat{f}_{N,M}(x) - f_M(x)]dx + \int_0^b [f_M(x) - f(x)]dx,$$

where

$$f_M(x) = \frac{1}{2\pi}\int_{-M}^{M} g(t,x)dt.$$

Examine these terms separately:

$$\begin{aligned}\left|\int_0^b [\hat{f}_{N,M}(x) - f_M(x)]dx\right| &\leq \int_0^b \left|\hat{f}_{N,M}(x) - f_M(x)\right|dx \\ &\leq b \cdot \sup_{0\leq x\leq b} |f_{N,M}(x) - f_M(x)| \\ &\leq \sup_{t\neq 0, x\in[0,b]} \left|\frac{\partial^2}{\partial t^2} g(t,x)\right| \frac{M^3}{N^2} \longrightarrow 0,\end{aligned}$$

since $M^3/N^2 \to 0$ and $(\partial^2/\partial t^2)g(t,x)$ is bounded by assumption. The point $t = 0$ is not included since it is always the endpoint of a sub-interval for N even. The points at which the derivative of $g(t,x)$ is evaluated are in the interior of these sub-intervals.

Now handle the next term:

$$\begin{aligned}\int_0^b [f_M(x) - f(x)]dx &= \frac{1}{2\pi}\int_0^b \int_{-M}^{M} g(t,x)dtdx - \int_0^b f(x)dx \\ &= \frac{1}{2\pi}\int_{-M}^{M}\int_0^b g(t,x)dxdt - \int_0^b f(x)dx.\end{aligned}$$

Billingsley (1986), page 356, proves that

$$\frac{1}{2\pi}\int_{-M}^{M}\int_0^b g(t,x)dxdt \to F(b) \text{ as } M \to \infty.$$

Thus

$$\int_0^b [f_M(x) - f(x)]dx \to 0 \text{ as } M \to \infty.$$

Thus the result $\hat{F}_{N,M}(b) - F(b) \to 0$ has been achieved.

Proof of Lemma 4.2.2

First find the moment generating function of $Y = (Z + d)'(Z + d)$, where $Z \sim N_Q(0, V)$ and $V = \text{diag}(v_1, \ldots, v_Q)$. Let $m_Y(t)$ be the moment generating

function. Then

$$
\begin{aligned}
m_Y(t) &= \int_{\mathbf{R}^Q} \frac{1}{(2\pi)^{Q/2}(\prod v_i)^{1/2}} \exp[t(z+d)'(z+d)] \exp[-z'V^{-1}z/2]dz \\
&= \int_{\mathbf{R}^Q} \frac{1}{(2\pi)^{Q/2}(\prod v_i)^{1/2}} \\
&\quad \cdot \exp\Big[-[z - 2(V^{-1} - 2tI)^{-1}td]'(V^{-1} - 2tI)[z - 2(v^{-1} - 2tI)^{-1}td]/2\Big] \\
&\quad \cdot \exp[td'd + 2t^2d'(V^{-1} - 2tI)^{-1}d] \\
&= \frac{\left|(V^{-1} - 2tI)^{-1}\right|^{1/2}}{(\prod v_i)^{1/2}} \exp\Big[d'[t + 2t^2(v^{-1} - 2tI)^{-1}]d\Big] \\
&\quad \cdot \int_{\mathbf{R}^Q} \frac{1}{(2\pi)^{Q/2}|V^{-1} - 2tI|^{1/2}} \exp\Big[-(z-a)'(V^{-1} - 2tI)(z-a)/2\Big]dz \\
&= \frac{\left|(V^{-1} - 2tI)^{-1}\right|^{1/2}}{(\prod v_i)^{1/2}} \exp\Big[d'[t + 2t^2(v^{-1} - 2tI)^{-1}]d\Big].
\end{aligned}
$$

Let

$$A = V^{-1} - 2tI = \operatorname{diag}(1/v_j - 2t),$$

which implies that

$$\frac{\left|(V^{-1} - 2tI)^{-1}\right|^{1/2}}{(\prod v_j)^{1/2}} = \prod_{j=1}^{Q} \Big(\frac{1}{1 - 2tv_j}\Big)^{1/2},$$

and

$$d'(t + 2t^2A^{-1})d = \sum_{j=1}^{Q} d_j^2 \Big(\frac{t}{1 - 2tv_j}\Big).$$

Formally, by substituting $t = it$, one finds the characteristic function to be

$$\phi_Y(t) = \prod_{j=1}^{Q} \Big(\frac{1}{1 - 2itv_j}\Big)^{1/2} \exp\Big[d_j^2 it/(1 - 2itv_j)\Big].$$

Identifying $v_j = \theta_j$ and $d_j = \sqrt{\theta_j}b_j$ yields the result.

VITA

William Pyle Alexander [redacted]

[redacted] He graduated from the University of Arkansas with a Bachelor of Science degree in Agricultural Economics in May of 1982. He was a senior scholar in his graduating class. He then went to Texas A&M to pursue a Ph.D. in Agricultural Economics. In December of 1985 he moved to the Department of Statistics. In February of 1989 he received the Connor Award. The Connor Award is given to the outstanding Ph.D. candidate in the Department of Statistics each year. William expects to receive his Ph.D. in May of 1989. He has accepted a tenure track position in the Department of Statistics at the University of Kentucky, effective August of 1989.

[redacted]